From DEPARTMENT OF CLINICAL NEUROSCIENCE
Karolinska Institutet, Stockholm, Sweden

GENETIC LANDSCAPE OF MULTIPLE SCLEROSIS SUSCEPTIBILITY BY LEVERAGING MULTI-OMICS DATA

Tojo James

Stockholm 2018

Published by Karolinska Institutet.
Printed by Eprint AB 2018

ISBN 978-91-7831-181-1

Genetic landscape of multiple sclerosis susceptibility by leveraging multi-omics data
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Tojo James

Principal Supervisor:
Professor Ingrid Kockum
Karolinska Institutet
Department of Clinical Neuroscience

Co-supervisor(s):
Dr. David Gomez-Cabrero
Karolinska Institutet
Department of Medicine

Associate Professor Maja Jagodic
Karolinska Institutet
Department of Clinical Neuroscience

Professor Tomas Olsson
Karolinska Institutet
Department of Clinical Neuroscience

Opponent:
Dr. Calliope Dendrou
University of Oxford
Nuffield Department of Medicine
Division of Wellcome Trust Centre for Human Genetics

Examination Board:
Associate Professor Carsten Daub
Karolinska Institutet
Department of Biosciences and Nutrition

Professor Ann-Christine Syvänen
Uppsala Universitet
Department of Medical Sciences, Molecular Medicine

Dr. Sven Nelander
Uppsala Universitet
Department of Immunology, Genetics and Pathology

To my family

ABSTRACT

The main objective of the research studies presented in this thesis is to study the genetic variants and the expression of genes that relate to Multiple Sclerosis (MS). MS is a polygenic disease with HLA-DRB1*15:01 allele as a strong risk factor. Currently there are more than 200 non-HLA regions identified for MS. However, most of the risk loci identified in those studies are primarily driven by the relapsing-remitting form of MS (RRMS). To identify risk factors specific for the primary progressive form of MS (PPMS) which is a smaller group of MS patients, we have examined the exomes of PPMS and RRMS patients matching to population based controls in a case-control study setting and reported risk variants and mutations that are associated to PPMS and RRMS.

The context of this study is during the 'post-GWAS' era, when researchers are primarily focused to understand the functional consequences of the genetic risk factors. Using the possibilities of transcriptomic and genotyping data, genes that correlate to the risk loci are identified in relevant cell types of MS. Several statistical methods are implemented to characterize the risk loci and replicate the findings in the context of disease. MicroRNAs (miRNAs), small non-coding RNAs which regulate gene expression at post-transcriptional level, have been identified dysregulated in autoimmune diseases, including MS. We used experimental autoimmune encephalomyelitis (EAE), a commonly used animal model for MS to understand the role of miRNA in the immune activation of EAE.

Next generation sequencing (NGS) methods were widely applied in all of these studies specifically at transcriptomic and genomic level of the disease. NGS methods are data intensive but have higher reliability. To test the reliability, we compared reported gene expression measurements for ostensibly similar tissue samples collected from different RNA-seq studies. We found an overall consistency on expression data obtained from different studies and identified the factors contributing to systematic differences. This thesis gives an overview of progresses happening in the area of MS genetics, EAE model for neuroinflammation and omics data analysis to address genetic regulation of disease.

LIST OF SCIENTIFIC PAPERS

I. **Next-generation sequencing identifies microRNAs that associate with pathogenic autoimmune neuroinflammation in rats.**
Petra Bergman, Tojo James, Lara Kular, Sabrina Ruhrmann, Tatiana Kramarova, Anders Kvist, Gordana Supic, Alan Gillett, Andor Pivarcsi and Maja Jagodic
Journal of Immunology 2013 Apr 15;190(8):4066-75

II. **Whole Exome Sequencing to Identify Genetic Variants Associated with Primary-Progressive Multiple Sclerosis**
Tojo James, Sahl Khalid Bedri, Paola Bronson, K.D. Nguyen, Karol Estrada, Aaron Day- Williams, Lars Alfredsson, Tomas Olsson, Anna Glaser, Jan Hillert, Ingrid Kockum
Manuscript

III. **Impact of genetic risk loci for multiple sclerosis on expression of proximal genes in patients**
Tojo James*, Magdalena Linden*, Hiromasa Morikawa, Sunjay Jude Fernandes, Sabrina Ruhrmann, Mikael Huss, Maya Brandi, Fredrik Piehl, Maja Jagodic, Jesper Tegner, Mohsen Khademi, Tomas Olsson*, David Gomez-Cabrero* and Ingrid Kockum*
Human Molecular Genetics 2018 Mar 1;27(5):912-928

IV. **Assessing the consistency of public human tissue RNA-seq data sets**
Frida Danielsson, Tojo James, David Gomez-Cabrero and Mikael Huss
Briefings in Bioinformatics. 2015 Nov;16(6):941-9

*shared authors

ADDITIONAL PUBLICATIONS

DNA methylation as a mediator of HLA-DRB1* 15: 01 and a protective variant in multiple sclerosis
Kular L, Liu Y, Ruhrmann S, Zheleznyakova G, Marabita F, Gomez-Cabrero D, James T, Ewing E, Lindén M, Górnikiewicz B, Aeinehband S, Stridh P, Link J, Andlauer TFM, Gasperi C, Wiendl H, Zipp F, Gold R, Tackenberg B, Weber F, Hemmer B, Strauch K, Heilmann-Heimbach S, Rawal R, Schminke U, Schmidt CO, Kacprowski T, Franke A, Laudes M, Dilthey AT, Celius EG, Søndergaard HB, Tegnér J, Harbo HF, Oturai AB, Olafsson S, Eggertsson HP, Halldorsson BV, Hjaltason H, Olafsson E, Jonsdottir I, Stefansson K, Olsson T, Piehl F, Ekström TJ, Kockum I, Feinberg AP, Jagodic M.
Nature Communication. 2018 Jun 19;9(1):2397. doi: 10.1038/s41467-018-04732-5.

Smoking induces DNA methylation changes in Multiple Sclerosis patients with exposure-response relationship.
Marabita F, Almgren M, Sjöholm LK, Kular L, Liu Y, James T, Kiss NB, Feinberg AP, Olsson T, Kockum I, Alfredsson L, Ekström TJ, Jagodic M.
Scientific Reporter. 2017 Nov 6;7(1):14589.

Sex influences eQTL effects of SLE and Sjögren's syndrome-associated genetic polymorphisms.
Lindén M, Ramírez Sepúlveda JI, James T, Thorlacius GE, Brauner S, Gómez-Cabrero D, Olsson T, Kockum I, Wahren-Herlenius M.
Biology of Sex Differences. 2017 Oct 25;8(1):34.

The Multiple Sclerosis Genomic Map: Role of peripheral immune cells and resident microglia in susceptibility
NA Patsopoulos, SE Baranzini, A Santaniello,P Shoostari,C Cotsapas, G Wong, AH Beecham, T James, J Replogle, I Vlachos, C McCabe, T Pers, A Brandes, C White, B Keenan, M Cimpean,P Winn,IP Panteliadis, A Robbins,TFM Andlauer, O Zarzycki,B Dubois,A Goris,H Bach Søndergaard, F Sellebjerg,P Soelberg Sorensen,H Ullum,L Wegner Thoerner,J Saarela,I Cournu-Rebeix,V Damotte,B Fontaine, LGuillot-Noel,M Lathrop,S Vukusik, A Berthele, V Biberacher, D Buck, C Gasperi, C Graetz, V Grummel, B Hemmer,M Hoshi,B Knier,T Korn, CM Lill, F Luessi, M Mühlau,F Zipp,E Dardiotis,C Agliardi, A Amoroso,N Barizzone, MD Benedetti,L Bernardinelli, P Cavalla, F Clarelli, G Comi, D Cusi,F Esposito, L Ferrè, D Galimberti,C Guaschino, MA Leone, V Martinelli,L Moiola, M Salvetti, M Sorosina, D Vecchio, A Zauli,S Santoro,M Zuccalà, J Mescheriakova, C van Duijn,SD Bos, EG Celius, A Spurkland,M Comabella, X Montalban,L Alfredsson, I Bomfim, D Gomez-Cabrero,J Hillert, M Jagodic, M Lindén, F Piehl, I Jelčić,R Martin,M Sospedra, A Baker,M Ban,C Hawkins, P Hysi,S Kalra, F Karpe, J Khadake,G Lachance,P Molyneux,M Neville, J Thorpe,E Bradshaw,SJ Caillier, P Calabresi,BAC Cree,A Cross,M Davis,PWI de Bakker, S Delgado,M Dembele, K Edwards, K Fitzgerald,IY Frohlich,PA Gourraud,JL Haines, H Hakonarson, D Kimbrough, N Isobe,I Konidari, E Lathi, MH Lee,T Li, D An, A Zimmer,A Lo,L Madireddy,CP Manrique, M Mitrovic,M, Olah,E Patrick,MA Pericak-Vance, L Piccio,C Schaefer,H

Weiner,K Lage,ANZgene, IIBDGC, WTCCC2, A Compston,D Hafler, HF Harbo, SL Hauser,G Stewart,S D'Alfonso, G Hadjigeorgiou,B Taylor,LF Barcellos,D Booth,R Hintzen,I Kockum,F Martinelli-Boneschi, JL McCauley,JR Oksenberg, A Oturai, S Sawcer, AJ Ivinson, T Olsson, PL De Jager
BioRxiv, 143933, 2017 (under revision in Science)

TGFβ regulates persistent neuroinflammation by controlling Th1 polarization and ROS production via monocyte-derived dendritic cells
Parsa R, Lund H, Tosevski I, Zhang XM, Malipiero U, Beckervordersandforth J, Merkler D, Prinz M, Gyllenberg A, James T, Warnecke A, Hillert J, Alfredsson L, Kockum I, Olsson T, Fontana A, Suter T, Harris RA
Glia. 2016 Nov;64(11):1925-37

Circulating miR-150 in CSF is a novel candidate biomarker for multiple sclerosis.
Bergman P, Piket E, Khademi M, James T, Brundin L, Olsson T, Piehl F, Jagodic M
Neurology Neuroimmunology & Neuroinflammation 2016 Apr 20;3(3):e219.

Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment.
Gustafsson M, Edström M, Gawel D, Nestor CE, Wang H, Zhang H, Barrenäs F, Tojo J, Kockum I, Olsson T, Serra-Musach J, Bonifaci N, Pujana MA, Ernerudh J, Benson M.
Genome Medicine. 2014 Feb 26;6(2):17

CONTENTS

LIST OF ABBREVIATIONS

ABF	Approximate Bayes factor
ASE	Allele Specific Expression
BBB	Blood-brain barrier
cDNA	Copy DNA
CIS	Clinically isolated syndrome
CNS	Central nervous system
CSF	Cerebrospinal fluid
DA	Dark Agouti
DNA	Deoxyribonucleic acid
EAE	Experimental autoimmune encephalomyelitis
EBV	Epstein-Barr virus
EDSS	Expanded Disability Status Scale
eQTL	Expression quantitative trait locus
FDR	False discovery rate
FPKM	Fragments Per Kilobase Million
GWAS	Genome-wide association study
GTEx	Genotype-Tissue Expression
HLA	Human leukocyte antigen
IL	Interleukin
iOND	Inflammatory other neurological diseases
kb	Kilobases
LCL	Lymphoblastic cell line
LD	Linkage disequilibrium
MHC	Major Histocompatibility Complex
MRI	Magnetic resonance imaging
mRNA	Messenger RNA
MS	Multiple sclerosis
NIND	Non-inflammatory Neurological Disease
NGS	Next generation sequencing
OCB	Oligoclonal bands
OND	Other neurological diseases
OR	Odds ratio
PBC	Population based controls
PBMC	Peripheral blood mononuclear cell
PCA	Principal component analysis
PCR	Polymerase chain reaction
PPMS	Primary progressive MS
PVG	Piebald Virol Glaxo
qRT-PCR	Quantitative reverse transcription PCR
QTL	Quantitative trait locus
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RPM	Reads Per Million

RPKM	Reads Per Kilobase Million
RRMS	Relapsing-remitting MS
SNP	Single nucleotide polymorphism
SPMS	Secondary progressive MS
TF	Transcription Factor
Ti/Tv	Transition to transversion ratio
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

1 INTRODUCTION

1.1 MULTIPLE SCLEROSIS

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS) leading to demyelination and neuronal loss which was first described in 1868 by a French neurologist Jean-Martin Charcot [1]. It is one of the leading cause of neurological disability among young adults worldwide, with a prevalence of 0.19% in Swedish population and a female to male incidence ratio of 2.26 [2]. MS is initiated by infiltration of immune cells across the blood brain barrier (BBB) leading to demyelination and neuronal loss with inflammatory lesions [3]. The variation in clinical manifestations in MS and different symptoms such as disturbances of motor function, sensation and vision depend to a large extend on the site of lesions within the CNS [4].

For diagnosis of MS based on McDonald criteria, two episodes of demyelinating attacks separated by time is required [5]. A first episode suggestive of MS is called Clinically Isolated Syndrome (CIS). To establish a diagnosis of MS, magnetic resonance imaging (MRI) to visualize the presence of brain lesion and expanded disability status scale (EDSS) help to determine the functional states of MS patients [5]. Around 90% of patients have oligoclonal bands and 70% of patients have increased levels of IgG in the Cerebrospinal fluid (CSF) [6,7]. About 85% of MS patients have relapsing remitting form of MS (RRMS) with repeated episodes of neurological symptoms. With time from disease onset, many of these patients accumulate neurological disability and progress to a more progressive form of MS called secondary progressive MS (SPMS) [8]. However, in cases of 10 to 15% of patients a progressive accumulation of disability from the disease onset are observed and are categorized as primary progressive MS (PPMS) (Figure 1). The current immunomodulatory treatments doesn't effectively halt progressive form of MS, however reduces the episodes of relapses [9].

MS is a complex disease in which genetic, environmental, epigenetic and life style factors determine the risk for disease susceptibility [10]. Many of these risk factors are found in healthy persons and moreover not all of these risk factors are present in any one MS patient, signifying the heterogeneous nature of the disease. To add to this complexity there is a component of genetic heterogeneity, implying that patients carrying similar clinical phenotypes can have different combination of risk genes for MS disease. In addition to risk genes, epigenetic factors and changes in gene expression or non-coding RNA can contribute towards developing this complex disease [11].

Although multiple factors influence the initiation of the disease, there exist a complex, multicellular pathophysiological process during disease progression of MS [3]. One of the generally accepted hypothesis is that MS is mediated by activated autoreactive myelin-specific T cells that infiltrate into the CNS, resulting in a chronic inflammatory response [8]. It is supported by genetic association of Human leukocyte antigen (HLA) class II molecules

and its presentation of CNS specific autoantigens to autoreactive T cells. Moreover, demyelinating lesions obtained from MS patients contain CD4+ and CD8+ T cells in addition to monocyte-derived macrophages and occasionally plasma cells [12]. There is also evidence suggesting a pathogenic role of autoreactive B cells [13]. After differentiating of B cells into plasma cells, it produce autoantibodies which are specific for myelin basic protein autoantibodies (MBP) and myelin oligodendrocyte glycoprotein (MOG) autoantibodies [14]. Moreover, patients with MS have immunoglobulin G oligoclonal bands (OCBs) in the form of B-cell activation in their cerebrospinal fluid (CSF) which is also a biomarkers used clinically for the diagnosis of MS [15]. Though the exact mechanism of B cells is unknown in MS, several anti-CD20 depleting drugs that target B cells such as Rituximab, Ocrelizumab and Ofatumumab have proven to be effective treatment of MS and provide more diverse and personalized treatment options for patients with MS [16].

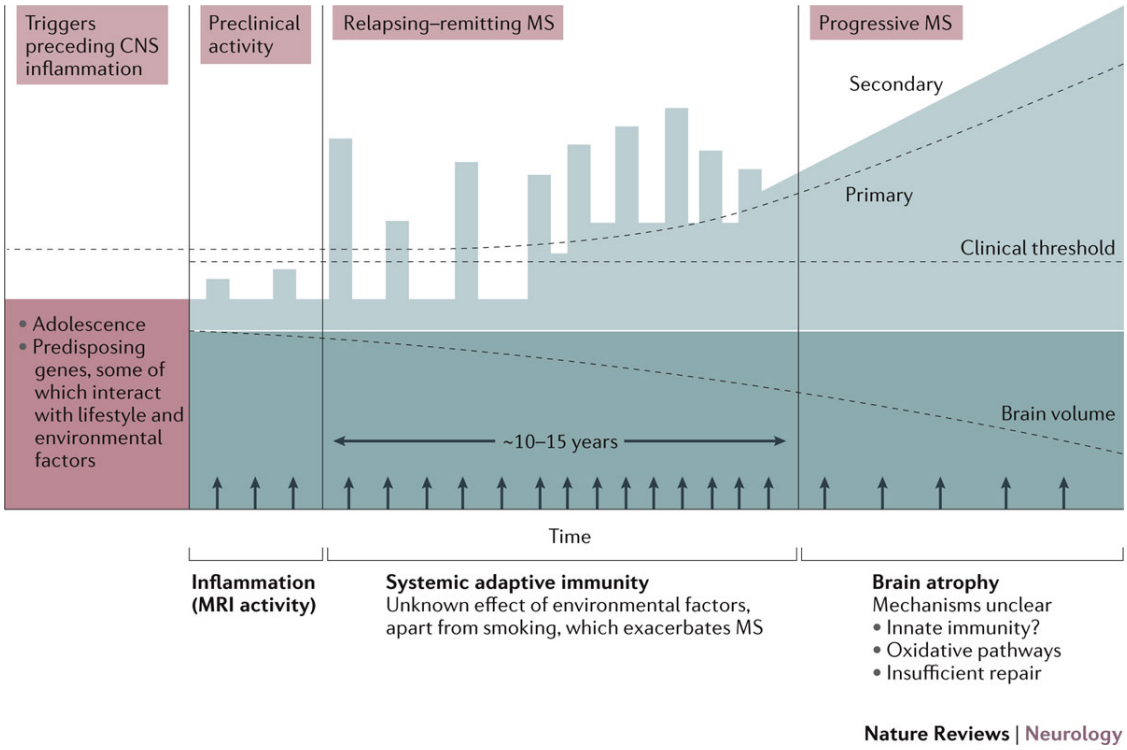


Figure 1: Progression and different phases of MS disease

[Figure reproduced from "Olsson, T. *et al.* (2016) Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis *Nat. Rev. Neurol.* doi:10.1038/nrneurol.2016.187"]

1.1.1 Genetics of Multiple Sclerosis

The human genome have more than three billion bases of DNA and its variation leads to genetic differences between individuals and populations. The common form of variation in the genome is the single nucleotide polymorphisms (SNPs) and imply a one-base at a fixed position in the genome [17]. A minor subset of SNPs correlates to the phenotypic differences including disease susceptibility and progression within and between populations. In addition to SNPs there are other classes of DNA variations that correlate to the observable phenotypes in the genome. These variations include large regions of variable copy number termed as

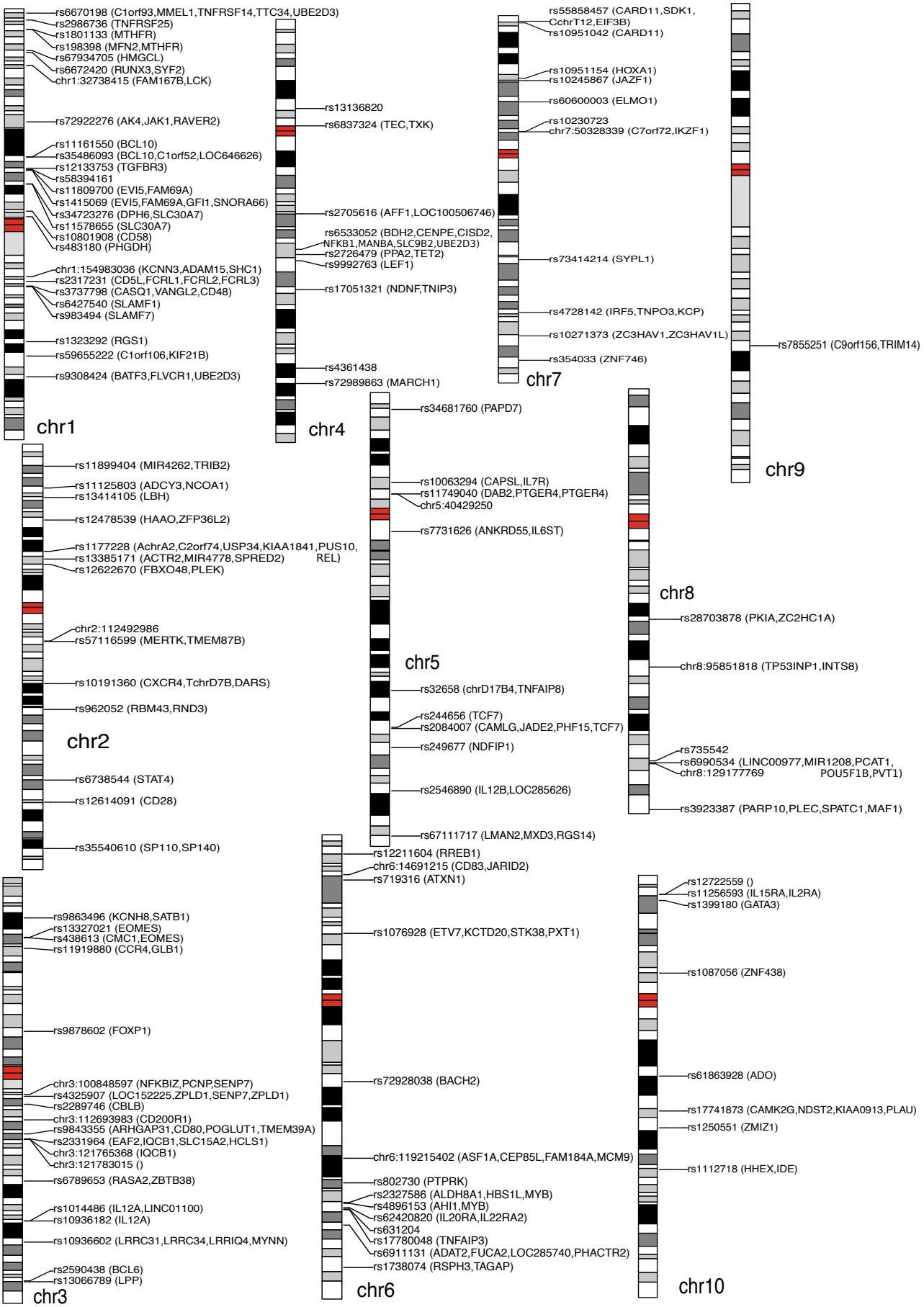
copy number variants (CNVs) and small variation or polymorphisms in the form of nucleotide Insertion and Deletion (INDELs). INDELs often appear to be in linkage disequilibrium (LD) with surrounding SNPs. These variants are widespread across the genome and can be found in both coding and non-coding sequences with varying frequencies across different populations [18]. If a variant found in the coding sequence alter the nucleotide triplets, called codons, resulting in an amino acid change of a protein molecule, the variant is defined as non-synonymous and conversely if it doesn't alter the amino acid sequence, it is defined as synonymous variant. Besides the direct alteration of protein through nucleotide change in the coding region, SNPs in non-coding region can vary the phenotype through transcriptional regulation or epigenetic control [19].

MS is a polygenic disease with one major risk loci in the HLA region. The genetic basis of MS was first demonstrated in familial aggregation studies and the overall proband-wise concordance rate for monozygotic twins was 18.4 which was significantly higher than for dizygotic twins at 4.6 and siblings at 2.7 [20]. However, a recent familial MS risk study based on Swedish population found a proband-wise concordance rate for dizygotic twins with same sex at 3.3 and observed a higher transmission rate of disease from fathers to sons compared to mothers to sons, suggesting the role of less prevalent sex in the disease transmission. The heritability estimated at 0.64 and the shared environmental component calculated to be 0.01, implies that genetics play an important role [21].

1.1.1.1 HLA and MS susceptibility

HLA genes located on chromosome 6p21 were the first reported genetic risk for MS which was identified through a hypothesis driven approach [22]. HLA gene complex encodes major histocompatibility complex (MHC) proteins in humans which have an important role in the regulation of immune system. MHC region in mammals are the most genetically variable coding loci and in humans there are currently 18,955 HLA and related alleles reported in the HLA database [23]. Based on its function, HLA genes are classified into three broad classes. HLA genes corresponding to MHC class I (A, B and C) are expressed in cell surface in most of the nucleated cells and can present self-peptides and intracellular pathogens including viruses to cytotoxic T-cells. HLA genes corresponding to MHC class II (DP, DM, DO, DQ, and DR) are expressed on the surface of antigen-presenting cells (APCs) and present peptides from extracellular components that can be degraded by the APCs. These particular antigens stimulate T-helper cells which in turn stimulate antibody-producing B-cells to produce antibodies to that specific antigen. HLA genes corresponding to MHC class III encode several components of the complement system (C2, C4a, C4b and Bf). Complement proteins have a role in activating and maintaining the inflammatory process of an immune response.

The HLA corresponding to class II allele variants — DRB1*15:01, DRB5*01:01, DQB1*06:02 and DQA1*01:02 show the strongest genetic association with MS. The DRB1*15:01 haplotype increases the MS risk by about 3-fold [24]. Although it has been fairly easy to identify MS associated haplotypes, it has been harder to identify which alleles on haplotype is responsible for the association. This is because of high LD in the HLA region. HLA corresponding to class I allele variants HLA-A*02 have a protective effect in MS [25].


rs6670198 (C1orf93,MMEL1,TNFRSF14,TTC34,UBE2D3)
rs2986736 (TNFRSF25)
rs1801133 (MTHFR)
rs198398 (MFN2,MTHFR)
rs67934705 (HMGCL)
rs6672420 (RUNX3,SYF2)
chr1:32738415 (FAM167B,LCK)
rs72922276 (AK4,JAK1,RAVER2)
rs11161550 (BCL10)
rs35486093 (BCL10,C1orf52,LOC646626)
rs12133753 (TGFBR3)
rs58394161
rs11809700 (EVI5,FAM69A)
rs1415069 (EVI5,FAM69A,GFI1,SNORA66)
rs34723276 (DPH6,SLC30A7)
rs11578655 (SLC30A7)
rs10801908 (CD58)
rs483180 (PHGDH)
chr1:154983036 (KCNN3,ADAM15,SHC1)
rs2317231 (CD5L,FCRL1,FCRL2,FCRL3)
rs3737798 (CASQ1,VANGL2,CD48)
rs6427540 (SLAMF1)
rs983494 (SLAMF7)
rs1323292 (RGS1)
rs59655222 (C1orf106,KIF21B)
rs9308424 (BATF3,FLVCR1,UBE2D3)
chr1
rs13136820
rs6837324 (TEC,TXK)
rs2705616 (AFF1,LOC100506746)
rs6533052 (BDH2,CENPE,CISD2,
NFKB1,MANBA,SLC9B2,UBE2D3)
rs2726479 (PPA2,TET2)
rs9992763 (LEF1)
rs17051321 (NDNF,TNIP3)
rs4361438
rs72989863 (MARCH1)
chr4
rs55858457 (CARD11,SDK1,
CchrT12,EIF3B)
rs10951042 (CARD11)
rs10951154 (HOXA1)
rs10245867 (JAZF1)
rs60600003 (ELMO1)
rs10230723
chr7:50328339 (C7orf72,IKZF1)
rs73414214 (SYPL1)
rs4728142 (IRF5,TNPO3,KCP)
rs10271373 (ZC3HAV1,ZC3HAV1L)
rs354033 (ZNF746)
chr7
rs7855251 (C9orf156,TRIM14)
chr9
rs11899404 (MIR4262,TRIB2)
rs11125803 (ADCY3,NCOA1)
rs13414105 (LBH)
rs12478539 (HAAO,ZFP36L2)
rs1177228 (AchrA2,C2orf74,USP34,KIAA1841,PUS10,
rs13385171 (ACTR2,MIR4778,SPRED2) REL)
rs12622670 (FBXO48,PLEK)
chr2:112492986
rs57116599 (MERTK,TMEM87B)
rs10191360 (CXCR4,TchrD7B,DARS)
rs962052 (RBM43,RND3)
chr2
rs6738544 (STAT4)
rs12614091 (CD28)
rs35540610 (SP110,SP140)
rs34681760 (PAPD7)
rs10063294 (CAPSL,IL7R)
rs11749040 (DAB2,PTGER4,PTGER4)
chr5:40429250
rs7731626 (ANKRD55,IL6ST)
chr5
rs32658 (chrD17B4,TNFAIP8)
rs244656 (TCF7)
rs2084007 (CAMLG,JADE2,PHF15,TCF7)
rs249677 (NDFIP1)
rs2546890 (IL12B,LOC285626)
rs67111717 (LMAN2,MXD3,RGS14)
chr8
rs28703878 (PKIA,ZC2HC1A)
chr8:95851818 (TP53INP1,INTS8)
rs735542
rs6990534 (LINC00977,MIR1208,PCAT1,
chr8:129177769 POU5F1B,PVT1)
rs3923387 (PARP10,PLEC,SPATC1,MAF1)
rs9863496 (KCNH8,SATB1)
rs13327021 (EOMES)
rs438613 (CMC1,EOMES)
rs11919880 (CCR4,GLB1)
rs9878602 (FOXP1)
chr3:100848597 (NFKBIZ,PCNP,SENP7)
rs4325907 (LOC152225,ZPLD1,SENP7,ZPLD1)
rs2289746 (CBLB)
chr3:112693983 (CD200R1)
rs9843355 (ARHGAP31,CD80,POGLUT1,TMEM39A)
rs2331964 (EAF2,IQCB1,SLC15A2,HCLS1)
chr3:121765368 (IQCB1)
chr3:121783015 ()
rs6789653 (RASA2,ZBTB38)
rs1014486 (IL12A,LINC01100)
rs10936182 (IL12A)
rs10936602 (LRRC31,LRRC34,LRRIQ4,MYNN)
rs2590438 (BCL6)
rs13066789 (LPP)
chr3
rs12211604 (RREB1)
chr6:14691215 (CD83,JARID2)
rs719316 (ATXN1)
rs1076928 (ETV7,KCTD20,STK38,PXT1)
rs72928038 (BACH2)
chr6:119215402 (ASF1A,CEP85L,FAM184A,MCM9)
rs802730 (PTPRK)
rs2327586 (ALDH8A1,HBS1L,MYB)
rs4896153 (AHI1,MYB)
rs62420820 (IL20RA,IL22RA2)
rs631204
rs17780048 (TNFAIP3)
rs6911131 (ADAT2,FUCA2,LOC285740,PHACTR2)
rs1738074 (RSPH3,TAGAP)
chr6
rs12722559 ()
rs11256593 (IL15RA,IL2RA)
rs1399180 (GATA3)
rs1087056 (ZNF438)
rs61863928 (ADO)
rs17741873 (CAMK2G,NDST2,KIAA0913,PLAU)
rs1250551 (ZMIZ1)
rs1112718 (HHEX,IDE)
chr10

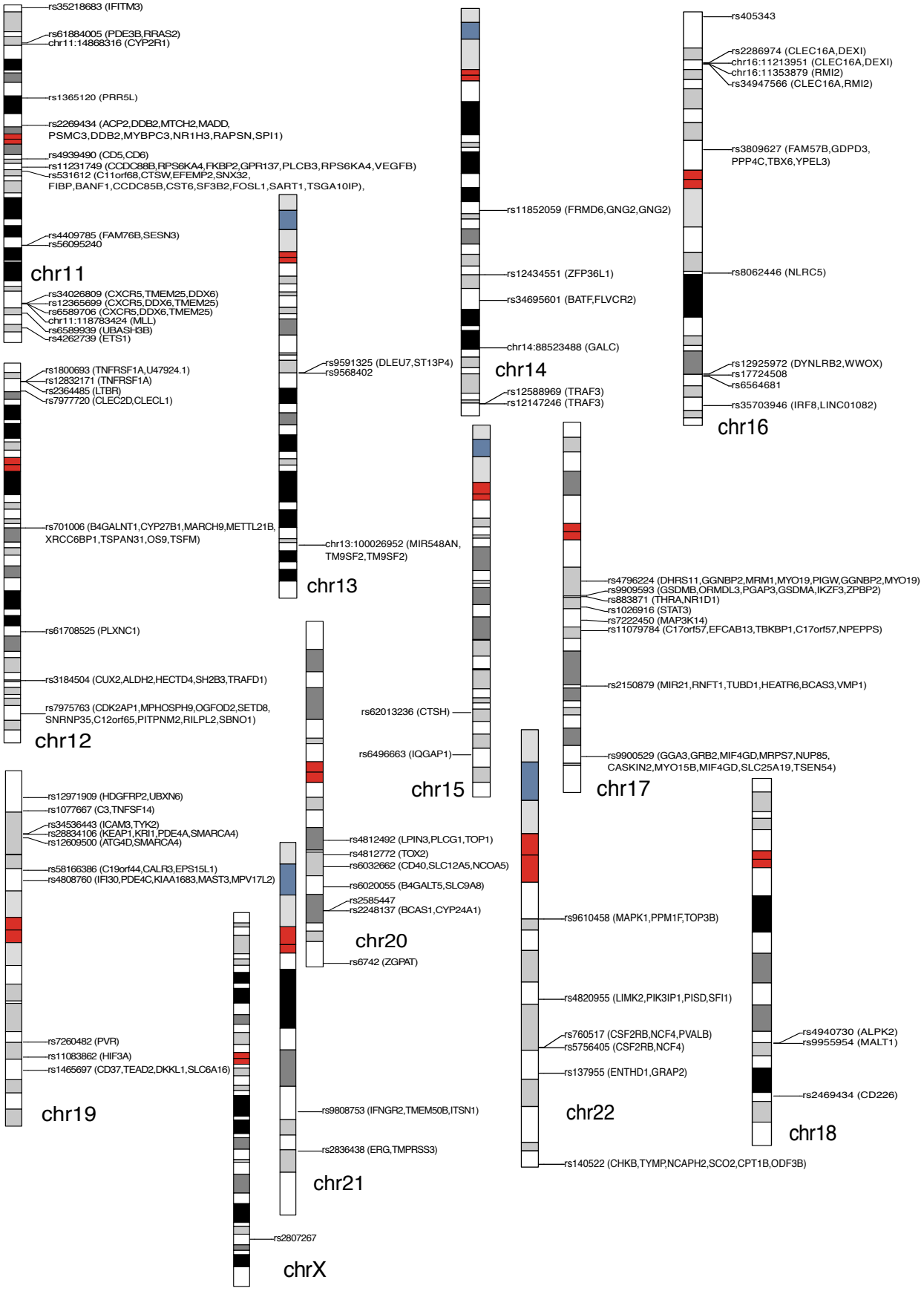


Figure 2: Association of 201 non-HLA SNPs (including chromosome X) on MS Chip study [33] and their corresponding eQTL genes. Red color in the cytogenic bands means telomere and blue color means centromere of the chromosome.

For a long time HLA has been considered as the only reproducible genetic risk factor in MS [26,27]. However, a recent study using a high-resolution mapping of HLA region has identified new associations in addition to the previous finding in class II risk alleles (HLA-DRB1*15:01, HLA-DRB1*13:03, HLA-DRB1*03:01, HLA-DRB1*08:01 and HLA-DQB1*03:02) and class I protective alleles (HLA-A*02:01, HLA-B*44:02, HLA-B*38:01 and HLA-B*55:01) [28].

1.1.1.2 Mapping MS risk genes based on large scale association studies

The landscape of genetics has changed remarkably by utilizing the possibilities of consortia based efforts, in particular genome wide association studies (GWAS) which include thousands of individual samples using new generation of genotyping platforms. The GWAS is a hypothesis-free approach in which SNPs tagging to the LD blocks across the whole genome are included on a genotyping array. Large sample sizes of cases and controls and measures for controlling for population stratification are important to perform well-powered GWAS.

The first MS-GWAS was performed by the International Multiple Sclerosis Genetics Consortium (IMSGC) using trio families, replicating in a case-control study and found two risk alleles outside the MHC region, mapping to the *IL7R* gene and *IL2RA* gene [29]. A second modestly powered GWAS for MS identified 29 novel non-HLA loci and three additional HLA loci and replicated 28 non-HLA loci and four HLA loci that were previously reported in different association studies [30]. GWAS chips contain only tag SNPs for common variants present in a population. To identify relevant SNPs that are associated to MS and to replicate the previous GWAS signals, ImmunoChip a custom-made genotyping array fine mapped on risk loci of eleven different autoimmune or inflammatory diseases was used [31]. The ImmunoChip study analysed 14,498 subjects with MS and 24,091 healthy controls for 161,311 autosomal variants and established 110 MS associated risk loci. Bayesian fine mapping was applied to refine these associations [32]. These large scale genetic studies resulted in the discovery of many non-HLA variants associated to MS disease. However the effect size of these non-HLA disease associated variants are smaller compared to the effect size of the variants that are strongly associated in the HLA region. A recent study (MS Chip study), analysed 47,351 multiple sclerosis (MS) subjects and 68,284 healthy subjects and reported 200 independent autosomal susceptibility variants including one variant in X chromosome and 32 independent associations within the HLA region [33]. In this study meta-analysis of several GWAS in MS was done in discovery part, followed by replication in independent cohort. Most of the loci found in this study are in close proximity to the immune-related genes which suggests them as potential candidate genes for involvement in pathogenesis of MS (Figure 2).

1.1.1.3 Missing heritability and rare variants

For many traits, large scale association studies have turned out to be highly successful with 24,218 unique SNP-trait associations from 2,518 publications reported in the NHGRI Catalog [34]. However these identified variants explain only a modest proportion of the total heritability. In genetics there has been much emphasis on so-called 'missing heritability' of traits. Some of the effects related to the missing heritability are contributed by epistatic (i.e. non-additive) interactions, so that the common variants contribute more risk in combination

than independently [35]. One of the other possible explanation is that these large scale association studies have only investigated common SNPs with a minor allele frequency (MAF) above 5% and lack proper knowledge of rare variants relevant in the disease which cannot be imputed by LD to the common variants present on the genotyping arrays. It has also been suggested that rare monogenic sub-phenotypes may exist for each common disease (Figure 3) [36].

A recent meta-analysis based on 32,367 MS cases and 36,012 controls analysed 144,209 low-frequency coding variants (MAF < 5%) genotyped using exomechip found four novel variants associated to MS independent of common variant signals found in previous studies (Figure 3). These novel genes have a key role for regulatory T cell homeostasis and regulation, IFNγ biology and NF-κB signaling in MS pathogenesis [37].

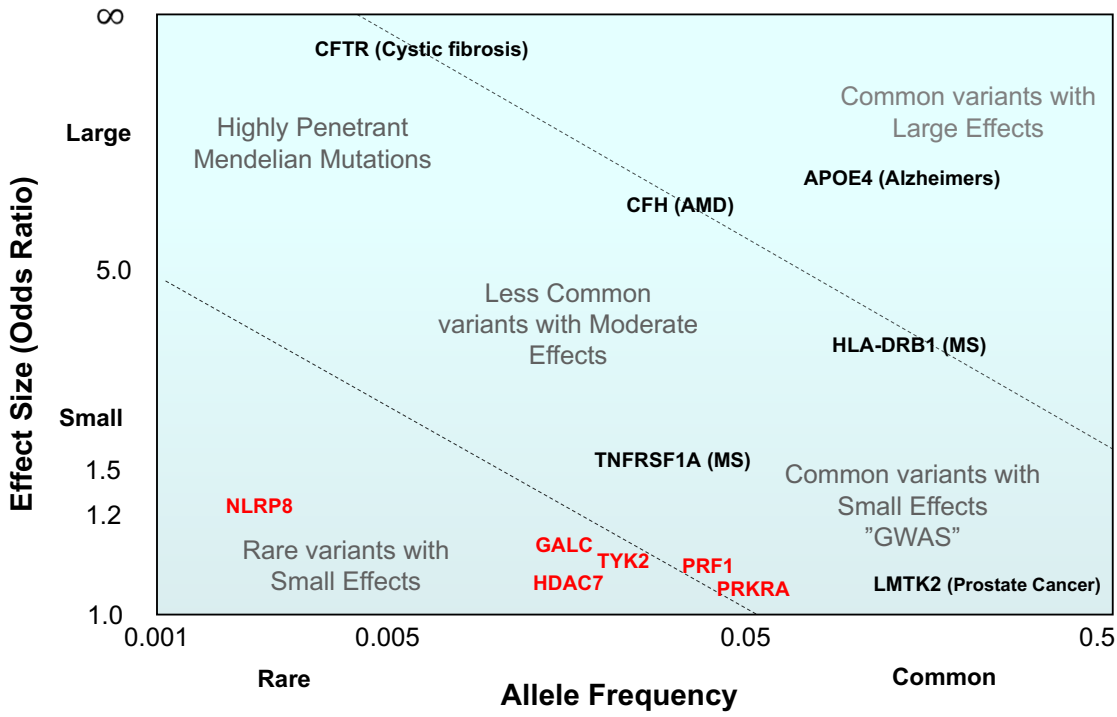


Figure 3: Spectrum of Disease Allele Effects. Scales in the axis are not linear. Genes marked in red are the genes with low-frequency coding variants studied in MS with exome chip [37].

1.1.2 Characterization of genotype-phenotype relationships in MS

Since many of the disease associated variants are located in the non-coding region of the genome and are synonymous or nonsense mutations (i.e. not changing the protein structure), the function and mechanisms of each of these variants are unknown [38]. This suggests that most of the common variants associated to the disease have regulatory functions [39]. Moreover, hundreds of genetic variants have been identified for many of the diseases, it is not feasible to knock-down or overexpress each of the genes within the risk loci. To identify the genes that are regulated by variants, Jansen and Nap in 2001 proposed the concept of "genetical genomics", by correlating variants to the intermediate molecular quantitative traits which

include gene expression, methylation or protein levels [40]. ***'Genetical genomics will combine the power of two different worlds in a way that is likely to become instrumental in the further unravelling of metabolic, regulatory and developmental pathways', Jansen RC and Nap JP (2001)*** **[41].**

1.1.2.1 From SNPs to genes — eQTL and ASE

Genes can be assigned to disease variants in different ways. A gene can be assigned to the SNP through its proximity or through a statistical or functional association with the gene. Since gene expression levels are strongly heritable (heritability in humans 0.25) and specific to tissues, it can be reliably mapped to the SNPs in different tissues [42]. Depending on the distance from the proximity of the SNP, eQTLs can act locally (cis) or at a distance (trans). cis-eQTL SNPs are located close to the transcription start site (TSS) of genes. However, for distance various measures have been used in different studies ranging from 5kb to 500 kb [43–45]. There is an increased probability of finding cis-regulated genes at closer distances to the variant and number of tests can be reduced by selecting fewer genes. This can also reduce multiple simultaneous statistical tests which is a drawback in trans-eQTL studies where the genes across different chromosomes are required to be tested (Figure 4). Moreover, trans-eQTL often requires very large sample size to reach statistical significance [43,46].

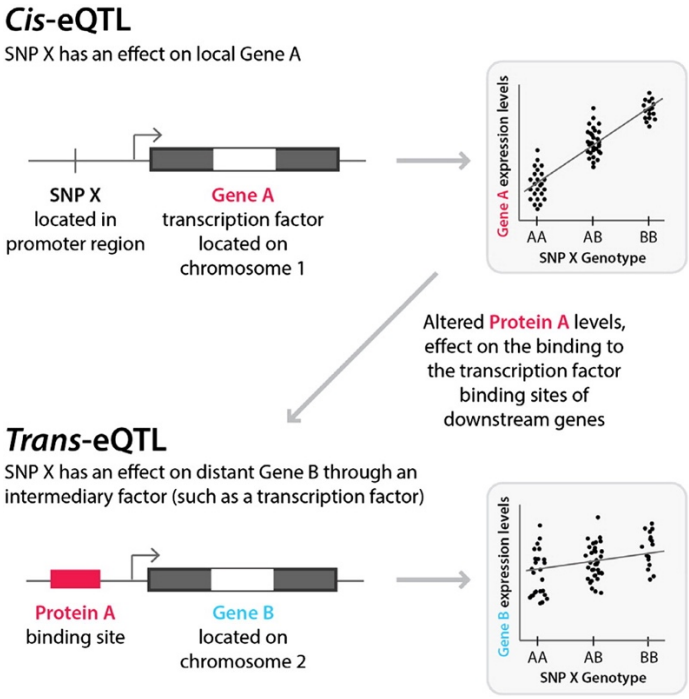


Figure 4: Possible mechanism of eQTL local effects (cis-eQTLs) and distant effects (trans-eQTL) [Figure reproduced from "Westra HJ, Franke L. From genome to function by studying eQTLs. Biochimica et Biophysica Acta - Molecular Basis of Disease. 2014. doi:10.1016/j.bbadis.2014.04.024"]

Cis-eQTL approach has its own fundamental limitations in disease study [47]. First, due to the LD it is challenging to identify the regulatory variant from neighbouring variants in moderate to high LD. Second, the expression of a gene can be influenced by multiple genetic, epigenetic and environmental or treatment factors. This suggests that eQTL effects may act in a context specific manner. Third, for a given variant with small effect size or with low allele frequency,

the study requires a large sample size to identify the effect. In study III of this thesis, different approaches and genetic statistical methods are implemented to address these concerns [48]. A complementary approach to find genetic variants associated to gene expression is based on allele-specific expression (ASE). In ASE, a regulatory variant which can be heterozygote can show different level of expression at gene level or transcript level [49]. Compared to analysing total expression levels in eQTL studies, in ASE the two alleles expressed at the same cellular environment and trans-acting environmental factors that can potentially increase variation between samples and individuals are minimized [50]. However, ASE is influenced by the local LD structure and level of allelic heterogeneity [47]. Therefore, these methods need to be implemented with caution where genomic loci are highly polymorphic, for example at HLA and T-cell receptor regions of the genome.

1.2 EXPERIMENTAL AUTOIMMUNE ENCEPHALOMYELITIS

Experimental autoimmune encephalomyelitis (EAE) is a widely used animal model to understand the mechanisms of inflammation in the CNS and role of the MS risk genes in vivo is studied in EAE models with the opportunities of gene targeting in mice [51]. Besides MHC locus in EAE, around 50 QTLs (Quantitative Trait Locus) have been reported to mediate EAE [52]. Some of the identified genes in both EAE and MS include genes in the MHC class II transactivator (*CLEC16A*) and *Il22ra2* [53–56].

To study the genetic regulation of neuroinflammation we have used two different inbred rat strains of EAE and recombinant MOG with incomplete Freund's adjuvant without pertussis toxin or mycobacterium is used to induce EAE [57]. In our study, to understand the regulation in miRNA in EAE (study I) we used the EAE-susceptible Dark Agouti (DA) rat strain and the Piebald Virol Glaxo (PVG) rat strain with identical MHC which is resistant to MOG induction. After the MOG induction an immune response in the secondary lymphoid organs is found in both strains but in DA rats the activation from myelin antigen become pathogenic in nature within 10-12 days. In DA rats the disease is characterized by demyelination in the CNS. An assessment for EAE scoring is done based on visual inspection of clinical symptoms such as ascending flaccid paralysis and measurements of weight loss.

The advantage of the animal model like EAE for neuroinflammation is the accessibility of appropriate tissues and target organ to study the disease pathogenicity and progression, which is difficult to obtain in human MS. In study I, we have collected the tissues from lymph node at different time points after the EAE immunization in DA and PVG rats. EAE is also characterized by lesions in CNS. However, compared to MS where lesions are found in brain which characterize the disease symptoms, in EAE the most affected site is the spinal cord. EAE model has provided an invaluable tool for development of new treatments against MS, including natalizumab [58] and glatiramer acetate and helped to understand the inflammatory process during the course of the disease [59]. However, there are certain therapies that had opposite outcomes in EAE compared to MS, for example administration of anti-tumour necrosis factor and interferon (IFN)-γ had adverse effect in MS patients [60].

In EAE myelin-specific T cells are activated in the periphery and infiltrate into the CNS through BBB [61–63]. In CNS the T cells are reactivated by infiltrated and local activated

antigen-presenting cells (APC), presenting major histocompatibility complex (MHC) class II associated peptides, leading to inflammation, demyelination and axonal damage. Studies from different models of EAE have shown evidence that T cells specific for self-antigens regulate the disease. Cytokine environment found in the draining lymph nodes during T cell receptor (TCR) stimulation is important for T cell differentiation. Until recently it was thought IFN-γ producing T helper type 1 (Th1) cells were the main effector T cells responsible for autoimmune inflammation [64]. Recent studies have highlighted the importance of IL-23 for EAE induction and autoimmune inflammation in human brain[65]. This led to the identification of IL-23-induced Th17 cells that produce IL-17, in addition to IL-17-secreting γδ T cells as a major effector cytokine [66,67]. Besides regulatory T cells, EAE was used to establish the role of B cells [68].

1.2.1 microRNAs in EAE and MS

microRNA (miRNA) is a small (about 22 nucleotides), single stranded non-coding RNA molecule which functions in RNA silencing and post-transcriptional regulation of gene expression[69]. Expression of more than 60% of human genes can be changed by miRNAs and similar to cytokines in immune regulation, miRNA function is characterised by pleiotropy where a single miRNA can target to many mRNAs and expression level of a mRNA can be regulated by many miRNAs, resulting in a complex regulatory network [70]. miRNAs are key regulators of the immune response, antibody secretion and release of inflammatory mediators and functional role in autoimmune diseases[71]. They have specific roles in both innate and adaptive immunity. A decline in the amount of Dicer or Drosha (essential enzymes for miRNA biogenesis) in regulatory T cells, can result in development of autoimmune diseases[72]. miRNA lacking CD4+ T cells failed to differentiate into Treg cells in the thymus. MIR155 knockout mice upon autoimmune inflammation were shown to be resistant to EAE due to reduced differentiation of Th1 and Th17 cells [73]. A recent study have also shown that expression of 56 miRNAs found in oligodendrocytes of EAE mice was lower compared to normal mice [74]. These increasing evidence shows that miRNAs are involved in various pathological conditions, including autoimmune inflammatory processes.

1.3 THE ERA OF OMICS

“Omics”, a suffix generally referring to the global measurement and analysis of a given level of biological information, encompasses the application of omics platforms range from genomics (identification of genes), transcriptomics (gene expression), epigenomics (epigenomic factors), proteomics (protein abundance), metabolomics (metabolites and metabolic networks) and pharmacogenomics (genetics effect on drug response)[75]. These data resources provide knowledge about molecular pathways in cells and their role in diseases. Presently, state-of-the-art next-generation sequencing (NGS) and multiplexed array technologies have enabled the large scale generation of experimental and clinical datasets [76]. Studies I to III of this thesis have utilized NGS methods within the scope of genomics and transcriptomics (including small RNA sequencing) to test various hypothesis related to MS and neuroinflammation in EAE.

Many large scale projects such as 1000 Genomes Project [77], ENCODE [78], ImmGen [79], TRANSFAC [80], Human Protein Atlas [81] etc have aimed to investigate biological systems at different levels generating large scale heterogeneous datasets with better annotation for the scientific community [82]. In all the studies included in this thesis we have utilized the potential publicly available datasets in an integrative manner to improve and refine the primary findings.

1.3.1 Genomics

Current array-based genotyping covers around 2.2 million SNPs, which is about ~2% of total common SNPs and have been widely used in many disease association studies including GWAS [83,84]. Two platforms— Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA), have been primarily used for GWAS. Illumina platform is based on a bead-based technology with a longer DNA sequences to detect alleles. Affymetrix have slightly shorter DNA sequences printed to a chip as a spot which detect a specific allele by differential hybridization of the sample DNA [85]. GWAS is based on Common Disease Common Variant hypothesis which implies that common diseases are regulated by the genetic variants that are common in population. If common variants are associated with a complex disease, the effect size (penetrance) must be small compared to that found in rare disease. Moreover, if the common variants have small effects (low penetrance) and if common diseases show heritability, then multiple common variants must regulate disease susceptibility. This suggests that the total genetic risk due to common variants spread across multiple genetic factors, prompting population-based studies in genome-wide scale compared to traditional family-based genetic studies [86]. Allele frequency of a variant and effect size of that potential disease variant are the key factors to be considered in addition to the sample size required to identify statistically significant genetic effects. Apart from common variants designed for GWAS these platforms also provide customized chips designed based on the region of interest (eg. ImmunoChip and MS Chip) or low-frequency variants (eg. Exomechip) [32,33,37]. In study III, for eQTL studies we have used genotyping data obtained from ImmunoChip and MS Chip.

Contrarily, Whole Genome Sequencing (WGS) method assays every nucleotide, resulting in several million variants including rare variants [87]. For cost advantage, low coverage sequencing (e.g. 4X average) is often preferred in WGS to maximize the cohort size. This increases the error rates for low coverage WGS to 15% or higher for discovery of variants [84]. As the cost of WGS become affordable, it will be the most promising technique to apply in large scale genetic studies [88]. Currently, compared to WGS the cost of Whole Exome Sequencing (WES) is lower and its promise is based on success of targeted gene sequencing studies and discovery of rare variants in the protein coding region with better coverage[89]. Since mutations are occurring in protein coding genes at a rate of $\sim 1 \times 10^{-5}$ per gene on a generation of nonsynonymous variants, every gene is expected to harbour functionally important variants which can be identified by sequencing. Moreover, potential genes underlying complex traits can be studied using WES and functional annotation of coding variants are usually simplified and straightforward in this approach [90].

1.3.2 Transcriptomics

To quantify the amount of mRNA in a sample several methods exist starting from northern blots and real-time PCR on single gene to microarrays and RNA sequencing (RNA-Seq) at global level. All these methods are different based on the range of gene expression measured and techniques used to detect the expression levels. Compared to other platforms, RNA-Seq does not depend on a pre-designed probe sequences, it lacks issues related to probe redundancy and annotation and thereby improves interpretation of the data. Though RNA sequencing achieved high resolution on the transcript architecture, the utility of other platforms is not undervalued and the selection of the platform purely depend on the study design and cost factor [91,92]. The application of PCR and RNA-Seq methods has expanded to include small RNAs and in study I, we have applied NGS methods to profile miRNAs and exon arrays to identify gene expression in EAE. Recent studies have shown that for high-resolution eQTL analysis, RNA-Seq can be considered as a 'gold standard', allowing a joint analysis of variation in gene expression levels and allele-specific expression across individuals [93,94]. In study III, we have applied RNA-seq to study the transcriptome of MS patients and disease eQTLs.

1.3.3 Omics data analysis and approaches

High-throughput 'omics' technologies enable the efficient generation of massive and complex datasets. Compared to the classical settings where very few specific null hypotheses are tested, the high-throughput nature of these technologies have resulted in the simultaneous analysis of very large number of variables (e.g. number of genes in expression or SNPs in array) compared to the number of independent subjects (e.g. clinical samples). This has resulted in simultaneous testing of many hypotheses, each subjected to a decision error (type I and type II errors), which requires careful adjustments in the form of multiple testing. To avoid overfitting and collinearity, classical statistical method requires the number of independent subjects to be large than the number of variables [95]. However, in omics based experiments the number of subjects are often limited due to economic and technical limitation of the experiments and therefore new statistical methods had to be applied specific to the experimental design and hypothesis under test. In study I to III, we have implemented different statistical methods in omics data analysis to test different hypothesis with better accuracy and significance.

In the exploratory phase of the analysis, dimension reduction and cluster detection approaches are commonly used in omics dataset [76,96]. While dimension reduction uncover the global variance within and between variables of dataset, cluster analysis explores the pairwise distance between the subjects for its relationships. These methods can explain the correlated structure of the data thereby identify the technical artefact such as batch effects or explain the characteristics of the outliers in the cluster analysis [97]. These methods are trivial while assessing the reproducibility, validity and interpretation of ostensibly similar omics data generated under different environments or stimuli. In study IV, we have used the possibilities of dimensional reduction and cluster detection to address reproducibility of the transcriptome data collected from different lab settings and platforms.

Depending on the omics platform and datatype the initial data processing steps and normalization procedures varies. Integration of different types of omics data in the context of a disease have a higher informative power compared to a single isolated omics data [98,99]. A single type of omics data can provide differences associated to a disease which can be reflecting either reactive process or causative factors [100]. To identify causative factors that lead to disease or to understand the disease mechanism, the integration aspect have a much wider scope as shown in this MS disease study [101]. In study III, we have applied a "genome first approach" where I assume that a disease associated variant contribute to disease rather than being the consequence of disease. These disease variants can be a direct source of risk prediction; however it may not suggest a particular gene or pathway that regulate the disease. To this end, a disease variant centered integration of additional omics data types such as transcriptome, DNASE etc, can imply the importance of the casual loci and pathways contributed to the disease.

2 THESIS AIMS

The general aim of the thesis was to study genetic risk factors for MS and to characterize its genetic regulation in humans and experimental model for neuroinflammation, utilizing next-generation sequencing methods.

Specific scientific goals:

Study I: To investigate different miRNAs regulated during pathogenic autoimmune neuroinflammation.

Study II: To identify different genetic variants associated with PPMS and RRMS

Study III: To gain further insights into the downstream effects of MS associated variants utilizing cis-eQTL based methods in patient derived primary cells and immune cells.

Study IV: To test the consistency among a set of publicly available RNA-Seq datasets obtained from similar tissues generated at different laboratories using different strategies.

3 METHODOLOGICAL CONSIDERATIONS

This thesis includes different bioinformatic methods, experimental techniques and cohort based dataset, which are described in detail in the included studies. Here I elaborate some of the relevant methods and approaches used for different studies in this thesis.

3.1 MIRNA PROFILE OF EAE AND ITS REGULATION

3.1.1 Small RNA sequencing and differential expression

After sequencing, small RNA sequence reads are demultiplexed to their corresponding samples and adapter sequence was trimmed from the read. In case of incomplete adapter sequence, a minimum adapter size of 8 bases was selected. The reads of length 12-42 bases were mapped to known miRNA of the miRBase repository (version 16.0) and miRNAs were quantified in terms of Reads per Million (RPM) with miRanalyzer tool using default settings [102,103]. For differential expression of miRNAs student t-test was applied and significant differentially expressed miRNAs were selected based on p value ($p < 0.05$) and expression values (RPM >1000). For low expressed miRNAs (RPM < 1000), fold change ($|FC| > 1.5$) was considered as an additional selection criteria.

3.1.2 miRNA target prediction and pathways in disease regulation

Limited knowledge on the mechanism of mRNA degradation by miRNA, poses development challenges for miRNA prediction tools. Genes regulated by differentially expressed miRNAs in EAE were studied using prediction tools. Several tools are developed considering the parameters such as seed sequence complementarity, site accessibility, sequence conservation, thermal stability of the miRNA:mRNA interaction etc, and to a wider extend these tools differ in their algorithms resulting in slightly different prediction. Considering these aspects, common gene targets predicted by two well established target prediction tools— TargetScan and miRanda were selected [104,105]. TargetScan is highly precise among the sequence-based tools, but have low sensitivity resulting from a high false negative rate. miRanda has better sensitivity but has a higher false positive rate [106]. To understand the regulation miRNA targets (genes) that are differentially expressed in EAE in the DA and PVG strains were selected for further functional analysis. This also improves the identification of relevant and true target mRNAs.

Using Ingenuity Pathways Analysis (IPA, https://www.qiagenbioinformatics.com/), relevant pathways regulated by differentially expressed genes targeted by miRNA were identified which can indicate downstream consequences of miRNA dysregulation in EAE.

3.2 WHOLE EXOME SEQUENCING STUDY ON MS PATIENTS

3.2.1 MS Cohorts for case-control study

The Genes and Environment in MS (GEMS) is a cohort based study in which a subset of cases in the Swedish Neuroregistry fulfilling the McDonald criteria were included besides prevalent cases of MS [107]. Epidemiological Investigations in Multiple Sclerosis (EIMS) is a population based case-control study of incident MS cases identified in neurology clinics throughout

Sweden [108]. IMSE cohort includes MS patients who were treated with Natalizumab [109]. In GEMS, EIMS and IMSE cohorts the controls are matched for age, sex, and residential area and are recruited from entire Sweden. For STOP-MS cohort, MS patients were recruited at the Karolinska University Hospitals based on Poser criteria or McDonald criteria for MS [110][5,111]. The controls include patients with other neurological diseases (ONDs) which can be either non-inflammatory or inflammatory. Subjects for study II were selected from the GEMS, EIMS, IMSE and STOP-MS cohorts. All PPMS samples included in the case-control study were matched to population based controls (PBC) based on age, gender and ethnicity. Gender and ethnicity were matched to the RRMS cases. Ethnicity information regarding individuals and their parents were obtained based on questionnaire data reported by the individuals.

3.2.2 SNP and CNV calling pipeline

In study II, genomic variants, such as SNPs and INDELs were identified using GATK v3.6 pipeline which primarily includes data pre-processing steps and variant discovery using variant calling and variant recalibrations algorithms [112]. The data pre-processing for GATK pipeline steps include mapping of sequence reads using BWA-MEM alignment tools and base recalibrations [113]. Gaussian mixture model (GMM) implemented in variant quality score recalibration (VQSR) step of GATK pipeline is applied to compare the quality of the variants in this study to highly validated variant resources (omni, 1000 Genomes, hapmap and dbSNP), thereby estimating the sensitivity and specificity for the variant calls. An optimal Transition to Transversion (Ti/Tv) ratio for exome sequencing is used to remove the false-positive variant calls in VQSR step [114]. A total of 1708 subjects were retained after exome mapping and variant calling step.

For copy number variants (CNVs) calling from exome sequencing reads we used CLAMMS [115]. For base-level depth-of-coverage calculation, a higher mapping quality and exon window coverage distributions were normalized based on overall sequencing depth and GC content. For a given sample CLAMMS compares its coverage data to probability distributions which describe the expected depth of coverage depending on copy number state, at each calling window. A total of 1666 subjects were retained, after subjects with inflated CNV rates and systematic biases in coverage data that fall under the outlier category of dimensional reduction technique (PCA) were filtered out.

3.2.3 Hardy–Weinberg equilibrium and quality control

Statistical tests for Hardy–Weinberg equilibrium (HWE) is an important method to detect genotyping errors which has been widely applied in population based genetic association studies. According to HWE principle, genotypes aa, aA and AA of a bi-allelic variant occur with relative frequencies of p^2, 2pq and q^2, where p and q are the allele frequencies for a and A alleles. HWE analysis aims to detect significant deviations from the expected proportions. Similar to SNP arrays, NGS methods are also prone to genotyping error specially in the low-coverage sequencing region and within the polymorphic locus of the genome which can deviate Ti/Tv ratio [87,116]. These problematic genotypes can affect the data quality for association studies. In Study II, we have applied HWE criteria implemented in vcftools for all the markers at a p-value $<10^{-5}$.

3.2.4 Association statistics for common variants

Single variant association tests were performed for common variants (MAF>5%) using logistic regression based on Wald test implemented in the EPACTS v3.3.0. Population substructure and unknown relatedness among the individuals in the cohort can have confounding effect on association studies [117]. Therefore, besides sex as a covariate twenty principal components (PCs) obtained from the kinship matrix which correct for relatedness was used in the model. The kinship matrix is obtained using vcf2kinship implemented in RVTEST [118]. In study II, in a case-control setting association tests were implemented for SNPs and INDELS in three study settings— PPMS vs PBC, RRMS vs PBC and PPMS vs RRMS and a p-value threshold of $5.5x10^{-7}$ is considered for significant association [119].

3.2.5 Annotation of relevant variants

In addition to annotating the SNPs with dbSNP names, the disease relevant variants were annotated with ClinVar [120,121]. This helps to find the rare variants or mutations that are important or closely related to the disease in the study. The variants with biological impact (deleterious or non-synonymous) were functionally annotated using Annovar database which is used in the burden test to access the biological impact of rare variants in a gene [122].

3.2.6 Association statistics for rare variants

In cases of large sample size with large effect sizes, single-variant tests can be used for rare variants but in study II, with limited sample size we applied methods that analyse variants jointly. Joint analysis of rare variants in a defined region requires statistical tests that are profoundly different from association statistics used for testing common variants [123]. Since exome sequencing captures the variations that can be annotated at gene level, the functional and population genetics information can be integrated into the test. Furthermore, rare variants need to be combined at gene level or in a specific pathway to reach sufficient power. In study II, we included rare variants which are non-synonymous (changing protein codon) and have high biological impact [124]. Broadly, based on summarizing the variants in the gene, these methods can be classified into three— Burden tests, Variance Component tests and Combination tests [125].

Burden test summarizes all genotypes in a gene into one collapsed genetic score which can be applied directly in the association test. Combined and Multivariate Collapsing (CMC) test is an extended form of burden test and it is used in study II. It "collapses" all rare variants in the gene region based on its minor allele and "combines" with the remaining common variants in the gene region in a “multivariate” problem setting which can be tested for a trait for that the gene region [126]. When analysing large gene regions with many variants or with missing genotypes resulting from the sequencing studies with low coverage, the burden scores tend to underperform. Cases and controls with differential missing data can account for an increase in type 1 error [125]. On the other hand, variance component tests such as SKAT, allow a mixture of effects in the given set of rare variants [127]. They assume that selected rare variants can either increase or decrease the risk of disease and tests are designed to account for varying effect sizes. Combination tests such as SKAT-O applied in study II combine burden scores with variance component [128]. There are different procedures to combine these two statistics. In SKAT-O, a convex combination of statistics with the weight coefficient is applied which

can be illustrated as a test statistic given by $Z_O = \alpha Z_S + (1 - \alpha)Z_B$ where Z_S is a SKAT based variance component test statistic and Z_B is a burden-like test statistic and the optimal weight coefficient α can be obtained by computing p values over a range of α values ($0 \leq \alpha \leq 1$).

3.3 MS TRANSCRIPTOME AND EQTL MAPPING

3.3.1 Patient cohorts and RNA sequencing

For the screening-phase of the eQTL study, RNA and DNA from PBMCs of 181 subjects were collected between 2001 and 2010, at the Neurology Clinic of the Karolinska University Hospital, Solna, Sweden (STOP-MS cohort). This cohort consisted of 117 MS patients, 28 CIS patients and 36 NIND (neurological diseases without an inflammatory state) patients. The reads from the RNA sequencing were mapped to hg19 genome (NCBI v37) using STAR aligner and quantified as counts per gene using HTSeq tool [129,130] To normalize the gene counts and to correct for the GC content and length biases, conditional quantile normalization (CQN) method was applied [131].

A separate cohort for validation with RNA and DNA isolated from sorted CD4+ and CD8+ cells were collected at the Neurology Clinic of the Karolinska University. Detailed report of the cohorts used in the screening phase and validation phase of this study is shown in Table 3 of the paper III.

3.3.2 Statistical framework for cis-eQTL analysis

For each disease associated SNP, we selected genes that were expressed in 85% of samples and located 400-kb upstream and downstream of the SNP. The window size was decided based on the increased probability of finding cis-regulated genes at closer distances and to limit the number of SNP-gene tests, with an aim to reduce number of statistical tests performed. To calculate the strength and significance of the SNP-gene association, we implemented a two-level analysis. First, the effect of selected variable identified in the component-based analysis is regressed out from the expression data normalized with Conditional Quantile Normalization (CQN) procedure [131]. Batch of RNA-Seq library preparation, age of individual at sampling, sex, disease-type, clinical course of MS and CIS and Interferon treatment were the variables in the meta-information investigated in the component-based analysis. Component-based analysis is explained in the later part of the methods under the section 'Component based analysis and Analysis of variance'. Second, spearman correlation is used to correlate residual values (obtained from regression) with genotype information obtained in the form of genetic additive model. In the additive model, genotypes are encoded numerically as 0,1 and 2. In case of a setting aa-aA-AA we used 0 for aa, 1 for aA, and 2 for AA, where allele 'A' is the risk allele associated to the disease. Permutation based p-value, with 15000 iterations was used to estimate the significance of the SNP-gene correlations. Non-parametric method was used to calculate False discovery rate (FDR) was implemented as described in [132] and to obtain sequential order as observed in Benjamini–Hochberg (BH) procedure, monotonicity is enforced in FDR estimation. A conservative selection criteria to reduce the false-positive findings; FDR<0.01 were considered for significant eQTLs. The eQTL study is extended to the LD SNPs ($r^2 > 0.5$) in the genomic locus of the association. A stepwise association model was applied where we

regressed out the effect of most-significant LD SNPs, obtained from the LD SNP-gene correlation, along with other cofounders such as batch and disease-type from the normalized expression levels (CQN values). LD SNPs with a beta-value greater than the original MS SNP are reported and the functional annotation (DNASE I hypersensitivity, Enhancer specific marks and TF motifs) of these LD SNPs for different immune cells are further characterized for this region. Custom R scripts and python programs were written for eQTL analysis.

3.3.3 Bayesian method for colocalization analysis

Colocalization investigates if a single variant in a locus is responsible for both disease genetic association signal and eQTL signal. An approximate Bayes factor was used to compute the posterior probabilities for variants that are considered causal in both disease and eQTL analysis. The R package "coloc" is used in the analysis to calculate the posterior probabilities for independent signals (H3) and shared causal signal (H4) from both eQTL association and disease association in a given locus [133]. In figure 5, each trait (eQTL signal and disease association signal) is represented as a binary vector (0,1) at 8 distinct genomic positions (SNP positions) which are shared between the traits. The value of 1 means that the SNP is causally involved in disease or an eQTL, 0 that it is not. The first plot shows the case where only one — eQTL signal or disease signal have an association.

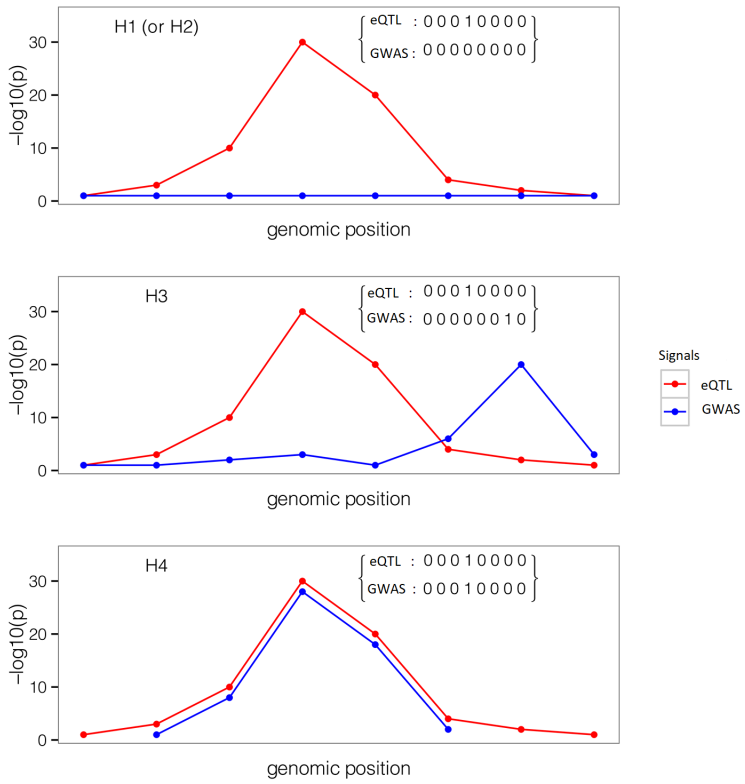


Figure 5: A configuration setting to illustrate different hypotheses.

[Figure adapted from "Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. *Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics.* PLoS Genet. 2014;10. doi:10.1371/journal.pgen.1004383"]

The second plot shows that the causal SNP is different for the disease association dataset compared to the expression dataset. The third plot shows the configuration where both signals

localize through a single causal variant (4th genomic position). The test was done in loci within 400-kb windows of MS-associated SNP and the prior probabilities (p1/p2 and p12) were set to the default values (p1/p2 = $1x10^{-4}$ and p12 =$1x10^{-5}$). Approximate Bayes factor (ABF) was computed for H3 and H4 using "coloc.abf" function. Spearman correlated p-values obtained from the eQTL analysis and association P-values of all the SNPs in the disease locus were obtained from the immunochip study. Only regions with a minimum of 10 independent SNPs after LD pruning (LD, r^2 <0.8) were selected for this analysis and colocalization with eQTL SNPs were considered if H4.abf was found greater than H3.abf.

3.3.4 Cell-type deconvolution and eQTL effect change

An in silico cell-type deconvolution method called CIBERSORT was applied to CQN transformed expression values obtained from PBMCs, using the LM22 signature geneset (1000 iterations) [134]. CIBERSORT allowed the estimation of the cell composition of PBMC from its PBMC gene expression profile using machine learning approaches. Cell types with a proportion >5% of the total PBMC mixture were aggregated into 5 groups: CD4 T cells, CD8 T cells, Monocytes, B cells and NK cells. Correlation between estimated cell-type proportions and disease associated SNPs and HLA variants was estimated; those correaltions with a p-value <0.05 were considered significant. For variants/SNPs with significant association, the effect of the corresponding cell type along with RNA-Seq batch covariate was regressed out from the gene expression levels (CQN values); then, on the residual values, eQTL analysis was conducted to test if , after correcting for cell proportions, there was still a SNP-gene significant association. A stepwise association model was implemented for those SNPs associated with more than one cell type; in those cases cell-type proportions were added as covariates in the order of cell type-SNP association strength.

3.3.5 Allele specific expression and analysis

Unlike other methods in ASE, where ASE is applied at single heterozygous variants, here we implemented phASER tool where it measures ASE at haplotype level [135,136]. This method use RNA-Seq data in combination with genomic data to phase heterozygous variants close to one another and aggregate the expression at the level of haplotypes. In our study we used genotyping data and PBMC RNA-Seq data used for eQTL study from 79 samples. Detailed steps of implementing genotyping data phasing and imputing procedures and ASE quantification steps are available here https://github.com/tojojames/PhaseImpute

3.4 RNA SEQUENCING AND REPRODUCIBILITY

3.4.1 Public RNA-Seq study

Processed RNA-Seq data and raw FASTQ files were obtained from five studies— BodyMap, Evolution of Gene Expression, Human Protein Atlas, RNA-Seq Atlas and AltIso [137–141]. The human tissues include heart, kidney and brain (hypothalamus in the case of RNA-Seq Atlas). The meta-information included information about study source, tissue type, preparation (RNA extraction), layout (single or paired end), read length (paired-end set to 200) and total number of raw reads.

3.4.2 RNA-Seq data analysis

To study the effects of reprocessing raw sequence data, FASTQ files for human heart, kidney and brain samples from each of the five sources mentioned above were mapped to the human genome (GRCh37) using TopHat [142] and FPKM values were computed using Cufflinks [143]. To study the effects of log transformation and to reduce the effect of extreme outliers, we applied log transformation directly on the untransformed published FPKM/RPKM values. To correct systematic study-specific effects such as batch effects which can introduce biases in the expression data we applied ComBat function included in the sva R package [144].

3.4.3 Component based analysis and analysis of variance

To study the consistency of clustering of tissue samples rather than by study we applied exploratory data analysis methods such as Principal component analysis (PCA). PCA is a simple eigenvector-based multivariate analyses method and it uncovers internal structure and variance in the data. In this gene expression dataset combined from different tissues, genes with highest PCA loadings are often tissue specific in nature. For this study, we examined correlations between principal component scores and study-dependent experimental factors or metainformation. Scores from the principal components which explain most of the variance, in total of at least 70% in the data, were selected for the correlation to metainformation. This helps to identify the factors or variables that contribute most to the variability of the data and similar approaches are applied in Study III to identify variables that influence the PBMC expression data obtained from patients.

The advantage of Analysis of variance (ANOVA) is that it can measure variability in a quantitative variable such as gene expression and partition it into various identifiable sources from a well-designed experiment. Primary sources of variability can include the experimental factors and random noise [145]. Since order matters in ANOVA test, we predefined an order that first includes the preparation variability such as layout, read length, RNA extraction method and number of raw reads and extended to the variability that cannot be controlled such as the study source and tissue types.

4 RESULTS AND DISCUSSION

The work presented in this thesis mainly covers multiple areas of genetic, transcriptomic and post transcriptional regulation of miRNA in the disease susceptibility and progression of MS.

4.1 MIRNAS IN POST-TRANSCRIPTIONAL REGULATION OF EAE (STUDY I)

After immunization with MOG on 5 susceptible DA and 5 EAE resistant PVG rats, miRNAs were evaluated from lymph nodes seven days after immunization using small RNA sequencing. From a total of 329 miRNAs evaluated, 43 miRNAs were found to differ between pathogenic and resolving immune activation of EAE and majority of differentially expressed miRNAs (35/43) were upregulated in the EAE-susceptible DA strain. Using specific TaqMan qRT-PCR assays for miRNA , some of the differentially expressed miRNAs were validated. To identify relevant genes in the EAE regulation, we combined predicted miRNA targets with expression data from lymph node of DA and PVG rats, collected seven days after immunization. With this integrated approach, we found 109 genes targeted by the upregulated miRNAs in DA strain and 54 genes targeted by miRNAs upregulated in PVG strain. Additional in vitro validation experiments based on Luciferase reporter system were done for rno-miR-181a target genes such as *Cxcr3, Prkcd*, and *Stat1* which have previously been reported in studies of MS and EAE. For highly expressed miRNAs, kinetics of its expression during the course of EAE was examined for a selected number of high-abundance miRNAs. For rno-miR-181a, rno-miR-128, and rno-miR-146a, the differential expression was observed at different time points post immunization (day 0, day 3, day 7 and day 25) and these miRNAs were primarily expressed in T cells. However, for rno-miR-223 and rno-miR-125b-5p differential expression is only observed on day 7 post immunization and these miRNAs were only expressed in non-lymphocyte cell populations.

From different expression kinetics of miRNA expression, it was determined that miRNAs have a role in regulating the ongoing inflammation and promoting disease. However, the role of individual miRNAs impacting the immune system and disease deserves further investigation. Using luciferase reporter system, we also validated three miRNA targets predicted for miR-181a, indicating that the targets we identify are not likely false positives. Since most of the miRNAs are conserved between human and rats, this study can provide some insights to the early development stages of MS.

4.2 GENETIC VARIANTS ASSOCIATED TO PPMS AND RRMS (STUDY II)

We explored potential genetic variants associated to MS using whole exome sequencing in cases and controls from PPMS, RRMS patients and population-based controls (PBC). With a cohort size of 552 PPMS, 575 RRMS and 575 age matched PBC; an association study was conducted at three levels—PPMS vs PBC, RRMS vs PBC and PPMS vs RRMS.

We found less significant association of HLA class I in PPMS patients compared to RRMS. However, for both PPMS and RRMS we found strong association to HLA class II as previously reported in MS studies and the significance disappear when correcting for DRB1*15:01 as a covariate in the model. In addition to previously reported associations in MS — *PHACTR2,*

P2RX7 and PRKRA we report novel suggestive associations in *CCL25, JMJD1C, LYZL2 and TIMM44* genes for RRMS and a significant indel association for PPMS in RP11-693J15.3 locus and two non-synonymous suggestive associations in *VCAN* gene [37,146,147]. From the Copy Number Variants (CNVs) association analysis, we find *LCE3C* deletion less common in PPMS patients compared to PBC and RRMS. In rare variant association studies based on burden tests, we found genes showing trend of association— *INO80D* associated to RRMS and *GBP5* associated to PPMS when compared to RRMS. Besides association studies, we also report the clinically relevant mutations observed in PPMS and RRMS cohort, which have resemblance to neurological disorders and MS.

Deletion of CNV in *PRKRA* hase been reported to be associated with ankylosing spondylitis risk and *PRKRA* activates PRK which regulates the antiviral effects of interferon [148,149]. Pro-inflammatory effect of the *P2X7* receptor that has been shown in various autoimmune diseases and is expressed on cells of the nervous system [150]. *CCL25*, a chemokine that signals through the receptor *CCR9* and is mainly expressed in small intestine and thymus, which have a key role in T cell maturation and recruitment of T cells to the small intestine [151,152]. *GBP5* stimulates *NRLP3* inflammasome assembly which initiates an inflammatory response towards pathogens and tissue injuries [153]. *NRLP3* has a role in the development of EAE [154]. *LCE3C* deletion is commonly found in the general population and deletion of *LCE3C* and *LCE3B* have been found to be associated to rheumatoid arthritis and psoriasis.

Using a small cohort, we have studied the potential genetic variants associated to PPMS and RRMS patients using whole exome sequencing (WES). Since PPMS represent a small proportion of MS, genetic association to PPMS have not been adequately tested in the previous genetic association studies. In contrast to genotyping platforms, WES have better resolution at the coding region of the genome and have the possibility to identify the variants that are in low frequency [155]. In the gene region of *PRKRA* reported in the exome chip study we have found indels associated to RRMS patients [37]. Studies have proposed that a significant amount of high quality genotypes outside target regions of the exomes can be obtained from WES data and in our study some of suggestive associations are outside the target region [156]. However this study was limited by sample size and require independent cohorts for replication.

4.3 DISEASE SPECIFIC EQTLS IN MS AND ITS RELEVANCE IN DIFFERENT CELL TYPES (STUDY III)

eQTL analysis in PBMC samples from MS and CIS patients were performed for 109 non-HLA and 7 HLA markers from Immunochip study and resulted in 77 significant eQTLs using RNA-Seq from PBMC samples (n=145). Permutation based methods for eQTL analysis were implemented after correcting for relevant covariates such as gender, diagnosis and batch of library preparation. Out of these some eQTL associated genes, for example *CPTIB, MANBA, PLEK, METTL21B, AHI1, TNFRSF14, MERTK, IQCB1* and *CLECL1* have been reported previously in studies from healthy individuals [44,157,158]. We also identified novel eQTLs; mostly non-coding RNAs, antisense transcripts and pseudogenes . "For 31 out of 47 non-HLA MS-eQTLs (66%) the transcription start site was located within 100 kb of the SNP, although 70.2% of the total eQTLs did not affect the gene closest to the SNP. We identified eQTLs associated with cell proportions — rs2726518 - *TET2* was associated with the proportions of monocytes, and rs6881706 -*SPEF2* was correlated with the proportions of B cells, CD8+ cells

and monocytes. Bayesian test for colocalisation of eQTL and genetic signal were performed in 22 MS susceptibility loci. 38 SNP-gene eQTLs were colocalised with the MS locus signal and two SNP-gene eQTLs — rs201202118-XRCC6BP1 and rs533646-AP002954.4.1, were not colocalised. Validation of MS eQTLs was perfomed in cell-type specific datasets- CD4, CD8 cells (RNA-Seq) and B cells and monocytes (microarray) from controls. Forty percentage of the non-HLA SNP-gene eQTLs found in PBMCs were significant in at least three additional cell types and 74% were significant in at least one cell type. MS eQTL effect changes in monocytes collected from healthy individuals activated by different stimuli (LPS, 2h and 24h; IFN-γ, 24h) was compared to the unstimulated situation. eQTL pairs such as rs8042861-IQGAP1 and rs941816-ETV7 showed increased effects in monocytes stimulated with IFN-γ and LPS (2h), rs2288904-SLC44A2 displayed an increased effect for IFN-γ, and rs2523822 (HLA-A*02)-HLA-H for each of the three stimuli, while rs11052877-CLECL1 showed a decreased eQTL effects after the three stimulations. Thirty two of 73 eQTLs had an effect size increase in MS cases compare to the pooled cohort of MS and NINDs (non-inflammatory neurological disease) and genes in HLA region showed the highest (eg. *HCP5*). *FCRL3* (eQTL gene), *FCRL2* and *FCRL5* genes in the rs706015 MS locus were significantly downregulated in MS cohort compared to NINDs. Mapping to enhancer histone marks and predicted transcription factor binding sites added additional functional evidence for eight eQTL regions. We identified that rs71624119, shared with three other autoimmune diseases and located in a primed enhancer (H3K4me1) with potential binding for STAT transcription factors, significantly associates with *ANKRD55* expression.

We tested and replicated the eQTLs found from PBMCs in datasets from CD4 and CD8 T cells of MS patients and healthy controls from LCLs, primary B cells and monocytes, including monocytes activated by IFN-γ and LPS. Since there are overlap in risk genes between autoimmune diseases, the results reported in this study are relevant for other autoimmune diseases. This study can be a resource of information for further functional exploration of genes in animal models of MS and to mine the signaling pathways that are affected by MS associated variants. In our study we found that eQTL SNP for *ANKRD55* was located in in an active regulatory region in PBMC and derived immune cells, with a possibility to bind to STAT1 transcription factor. *ANKRD55* was studied in EAE at protein expression level in neuron and microglia cultures, showing an active role in neuroinflammation and in humans the region is associated to Crohn's disease, Rheumatoid Arthritis ad Juvenile Idiopathic Arthritis [159,160].

4.4 CONSISTENCY OF PUBLIC HUMAN TISSUE RNA-SEQ DATASETS (STUDY IV)

In recent years, the accessibility of RNA-Seq based expression data have increased through different web repositories where users can get either precomputed gene expression values or raw sequence from various tissues collected from different organisms. Since universal standardization of RNA-Seq experimental protocols and equipment across laboratories is not practically possible, comprehensive studies of reproducibility focused on RNA-Seq data analysis pipeline are required. Furthermore, some studies have shown that publicly available RNA-Seq data sets from various laboratories and studies are at different phases not reproducible owing to the combined influence of reagent batch effects, differences in RNA extraction protocols, library preparation methods and computational processing [161–164].

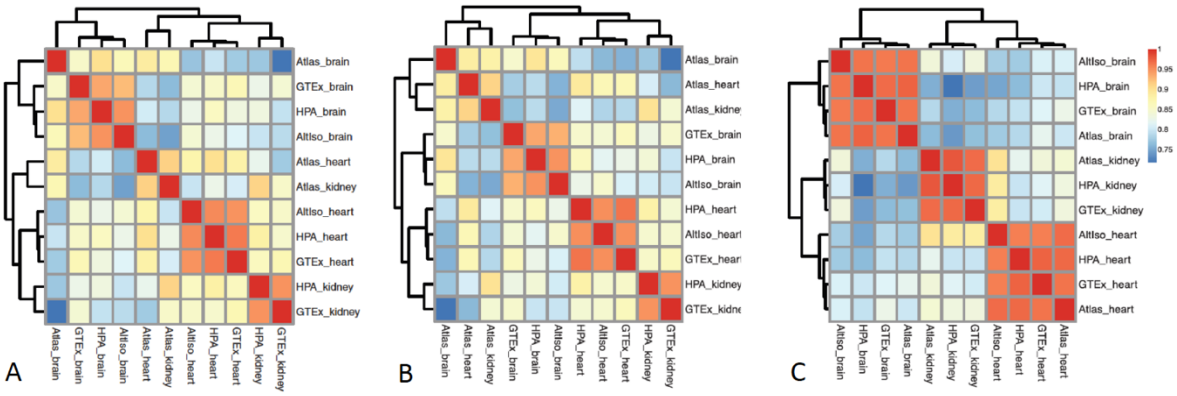


Figure 6: Heat map with Spearman correlations between samples from brain, kidney and heart (A) Analysis of 11 data sets with published precomputed FPKM/RPKM values. (B) Analysis of log-transformed published precomputed FPKM/RPKM values (C) Analysis of precomputed FPKM/RPKM values from four different studies after removal of batch effects using ComBat.

To gain further insight into the reproducibility issues, we evaluated published human tissue expression profiles from brain, heart and kidney by quantifying factors contributing to variability between samples, using analysis of variance (ANOVA) and principal component analysis (PCA). From our analysis we found that reported FPKM/RPKM gene expression values are not comparable at global level. The read length was the factor that contribute to the highest variance for the reported gene expression data sets followed by library preparation method, layout and tissue type (Figure 6). However after log transformation and removal of batch effects, the data show global consistency. Besides the advantage of not missing gene identifiers for various studies, there is no additional contribution towards clustering by simply reprocessing the raw RNA-Seq data obtained in FASTQ format. This study demonstrates the usefulness of public RNA-Seq data and provide a positive outlook on the continuous efforts to secure and precautious on application of datasets collected from different laboratories.

5 CONCLUSIONS AND FUTURE PERSPECTIVE

The underlying theme throughout this thesis has been that patient sample biobanks and omics based techniques are crucial in answering fundamental questions of disease biology and further improvements in this area can progress the future research in a faster pace. However in MS where the tissue for the disease is initiated and the processes leading to progression occur are not directly accessible in patients, it is important to use model organisms to understand the pathophysiology of the disease. Similar to all animal models in disease study, EAE has limitations; as it is heterogenous with respect to induction methods and pathological feature; and its utility depends on using the appropriate model to answer specific questions [174,175]. We have studied the role of miRNA in the epigenetic regulation of neuroinflammation within the setting of EAE. With the scope of NGS methods, we sequenced miRNAs from two inbred rat strains with higher quality and resolution which was superior to the array based or PCR based methods. One of the observation is that the dispersion in miRNA expression and the fold change of the top differentially expressed miRNAs is lower compared to gene expression. This might be due to feedback mechanisms that control miRNA levels or miRNAs might have a central role in maintaining the stability and plasticity of the immune system without directly initiating the disease. This study utilized the scope of prediction tools to identify miRNA targets. This was challenging as different tools predicted genes with slightly varying overlap on 43 miRNAs that we aimed to study. With the approach of integrating gene expression data to this study, we gained an added opportunity to select the genes that are both downregulated and targeted by miRNA, thereby helping to mine relevant pathways regulated by miRNAs in the experimental model of neuroinflammation. Using EAE, we also have shown that miRNAs have a crucial role during the early time-point of initiation and development of disease, which is not possible to study in human disease.

One of the objective of studying the risk genes of PPMS was to fill the gaps in our understanding of the genetics of MS which was predominantly guided by RRMS patients with a distinctive clinical manifestation and a higher incidence in females. PPMS being invariably much smaller compared to RRMS, little is known about risk factors associated with the onset of PPMS [167]. Our study is limited by the cohort size, which restricted our findings to risk variants with high effect size. However by investigating the exomes of PPMS and RRMS we found a less strong association of HLA class I molecules in PPMS compared to RRMS. There are studies in animal model showing that CD8+ T cells with restricted MHC class I are capable of initiating the first subtle attack [168]. With whole exome sequencing and improved variant annotation, it is possible to study the association of the rare variants in the disease. Since we are restricted by the sample size, different gene based tests are applied to understand the combined effect of rare variants of higher functional impact. These methods depend to an extend on the rare variant annotation, scoring methods used to establish protein altering SNPs (eg. CADD score), inclusion criteria for rare variants in gene burden test (eg. at least 2 rare variants of impact in a gene) and selection criteria based on allele frequency for rare variants. In addition to the gene based tests reported in the study II, we have applied different settings for gene based test. In one of the setting of SKAT-O test, genes with at least two rare (MAF < 0.005) stop gain, splice donor, splice acceptor, or missense variants were selected where variants were predicted to have a damaging effect in protein level by setting

the CADD score > 20 or PolyPhen2 parameter set to “probably damaging”. This helped to find interesting genes with significant association for PPMS.

Genome-wide scans, specially rare variant studies often are subject to a “winner’s curse” phenomenon resulting from the overestimation of genetic effects of associated variants and by underestimating sample sizes which can even result in opposite effects for the same variants in the replication study [169]. Since there are genome sequencing studies done in Swedish population, we can increase the sample size for population based healthy controls. To this end, for disease associated variants we have compared allele frequencies of healthy controls in our study cohort to allele frequency reported in SweGen project and found most of the reported variants having similar allele frequencies [170]. Moreover, for this study we have sequenced exomes with sufficient depth and followed GATK pipeline following specific parameters to improve the specificity and sensitivity of the variant discovery. This have resulted in finding mutations with high reproducibility and found the subject carrying the same mutations in *CSF1R* gene [171]. In one of the PPMS patients in this study, we found previously reported mutation NR1H3p.Arg415Gln, which was first identified and reported by Wang et al from the members of families with a progressive course of MS [172,173]. We have also scanned for mutations relevant for other neurodegenerative diseases in PPMS and currently we are examining their consequences in a clinical setting.

Understanding the genetic risk architecture for an associated set of variants is important to elucidate the causes or pathological mechanisms occurring in a multifactorial disease such as multiple sclerosis. This objective was broadly achieved by characterizing the genotype-phenotype relationships in MS through eQTL studies. With a high resolution transcriptomic data generated using NGS methods, applying appropriate statistical methods and integration of different cell type data we studied in depth the risk locus of 110 MS associated SNPs and 7 HLA variants reported in ImmunoChip study. We have reported many novel and validated eQTLs interesting for future functional characterization of MS. Previous study have shown that eQTL genes are enriched in a cell-type specific manner for different diseases and in case of MS it was enriched in CD4 cells compared to monocytes [44,165]. In our study we extended the eQTL study to four PBMC derived immune cells with different stimulations for monocytes after the initial screening in PBMC of patients. About 74% of the eQTLs were replicated in at least one of the immune cells. This suggest that although gene-regulatory regions bearing disease risk variants are accessible in multiple immune cells, it may regulate gene expression in either a cell-type specific or condition-specific manner. With Genotype-Tissue Expression (GTEx) project, our study can be further extended to more relevant tissues for disease such as brain. GTEx project report eQTLs based on bonferroni-corrected p-values which is a stringent criteria compared to the nominal p-value (p-value <0.05) criteria selected for replication part of our study [166]. We have also studied eQTL in the disease context by comparing MS cohort to pooled Cohort of MS and NINDs. Since cohort size of NINDs is small, these results need to be replicated and studied in larger cohorts. Two eQTLs were significant associated with cell proportions— rs2726518 associated with *TET2* was correlated to proportions of monocytes, and rs6881706 associated with *SPEF2* was correlated with the proportions of B cells, CD8 cells and monocytes. This is one of the first study where we investigated the effect of cell type proportions jointly with eQTL study. Similar to colocalization method applied in this study, it can be possible to model the SNP correlation

to cell types, gene expression and disease association and one can test the hypothesis that cell type proportion is leading the gene expression changes in PBMC. If cell type proportion have a leading role in regulating gene expression, one can expect cell-type specific methylation changes in these regions in the same primary tissue or in one of the derived cell types. With the epigenomic mapping of MS-associated variants, eight eQTL regions overlapped either with an H3K27ac or H3K4me1 active enhancer histone marks or with a DNase I hypersensitivity peak in monocytes, B cells, CD4 T cells and CD8 T cells and two eQTL regions overlapped with transcription factors. These epigenomic mapping provide some insights into complex regulatory mechanism localized to that particular region and help to prioritize appropriate cell type for translational studies.

These are the exciting days to work in the field of computational biology. With the ever growing collection of biological data and repositories, it is possible to test different hypothesis and develop methods in a quick turnaround time. We utilized the possibilities of applying previously published data in all our studies and in particular to study the viability and biases of applying tissue specific RNA-Seq data that are publicly available. The idea of tissue specific RNA-Seq study originated at a time when reproducibility aspects of expression data collected from different laboratories was questioned and from the observations that we made in our in-house RNA-Seq studies. Our original expectation was; to gain higher accuracy and reliability, one have to redo the initial mapping procedures of RNA-Seq analysis pipeline for every publicly available datasets. To our surprise with slight pre-processing and correcting for batch effects, it was possible to show that published human tissue RNA-Seq expression measurements are relatively consistent, suggesting the high reliability of RNA-Seq technology. These finding are been confirmed by recent studies and have extended to more recent datasets generated by GTEx, HPA and The Cancer Genome Atlas (TCGA) [176,177].

It has been a decade since the first GWAS in MS was published. Since then further efforts were directed towards identifying more independent risk variants using larger cohorts. In recent MS Chip study, suggestive SNPs which do not meet genome-wide significance ($p < 5\times10^{-8}$) contributed 9 % to the heritability so that now 48% of the heritability is estimated to have been explained [33]. Using WES, we have identified risk variants for RRMS and PPMS in suggestive and genome-wide significance level. Future association studies using large sample sizes, informed through WGS or WES may validate these new variants. One of the goal of association studies was to steer research toward novel genes involved in complex diseases. With increased number of identified risk loci, it is important to leverage relevant omics data to effectively prioritize and understand disease association signals and some of the approaches in the form of eQTL studies and WES (coding region of gene) are discussed in this thesis. eQTL studies can be extended to more relevant tissues, especially for the newly identified risk variants. Although we have shown the importance of eQTL studies in patient cohort, it would be beneficial to test this further in larger independent cohorts. We have studied the role of miRNAs in neuroinflammation and discussed challenges in understanding its role in disease regulation. Since miRNAs can target genes in a collaborative or competitive manner, it is difficult to attribute a miRNA-dependent dysregulation in disease and further studies in the form of gene (miRNA target) knock-down are required to completely understand the role of a selected miRNA in a disease.

6 ACKNOWLEDGEMENTS

Prof. Ingrid Kockum, I am lucky and grateful that I have had you as my main supervisor. You have a great scientific knowledge and it's amazing to see how you contribute to different projects within the group and among collaborators. Thank you for all your guidance, help and for the time you spend towards my PhD supervision and doctoral thesis. One of our project took longer time to finish than we thought and I am grateful for your patience and support. Thank you for the Julbord and parties at your place and for creating a wonderful group.

My co-supervisor **David Gomez-Cabrero**, Thank you for providing a nice "framework" for my PhD studies. I benefited a lot from your knowledge and guidance. Thank you for teaching me different statistical methods and for all the help with the eQTL paper. It's remarkable to see how quickly you answer to emails, even though you are working from different places. Thank you for welcoming me to your home at Valencia.

My co-supervisor **Maja Jagodic,** I came to CMM as your master thesis student. I guess that master thesis had an impact on me ☺. Later, you saw me enrolling as a PhD student. Thank you for all the inputs, you provided during my PhD studies and allowing me to participate in your projects. You are an excellent researcher with great motivation for science.

My co-supervisor **Prof. Tomas Olsson,** Thank you for allowing me to do my PhD studies in your unit. You have built a great research environment where everyone is enjoying their work. I am so amazed by your great knowledge in MS, passion for science and generosity. Thank you for giving me opportunities to attend international courses.

Prof. Jesper Tegner, thank you for the collaboration and inviting me to your group events. I learned a lot from the invited speaker talks that you hosted in your group. Now I am missing all those multidisciplinary talks.

Magdalena, thank you for sharing your knowledge about MS genetics specially at the beginning of my PhD studies. I don't know how many times we revised our eQTL paper to get it published. You are very inspirational in many ways and I was lucky to have you as my colleague. **Sahl,** finally we managed to run the exome sequencing project and thank you for your contribution towards the project. You are an amazing co-worker. I am happy to see you programming these days and thank you for all the late evening discussions.

Sreeni, thanks for your friendship and the long talks outside the lab. We started our PhD studies together and I am happy we are completing it together. For long time we were motivating each other to finish the PhD studies and finally we are very close to finish. Good luck with your preparations. **Sohel,** you are a very generous person and thank you for the advices on life and work balance. We still have to go for that long drive that we always plan for the summer.

Mikael Huss, I learned the basics of RNA-Seq analysis from you and it helped throughout my PhD studies.Please keep writing the blogs, I am following it. **Petra,** thank you for introducing me to the microRNA world. I really benefited from all those studies.

Pernilla stridh, thank you for all the help in getting the datasets on time for my projects. You have an in-depth knowledge in the areas you worked. Thank you for sharing it in our small

group meetings. **Sandra,** thank you for all the coordination efforts you are making in our group. We all benefited from it. Apart from coordinating, it's amazing to see that you are still helping with research work in our group. **Jesse,** you have so much energy and passion for science. I wish you all the best for your PhD! You deserve a bright future! **Katarina Tengvall,** thanks for your positive energy and introducing me to new methods in genetics. **Ali,** for the long discussion outside lab about MS and treatments. I learned a lot from you during a short time.All former colleagues **Magnus**, **Iza, Nada, Shahin, Emilie, Samina** and **Cecilia**, thank you for all the inputs inspite of your busy work, especially when I was starting my PhD studies.

Bob Harris, thanks for all the great work you contribute towards improving our PhD education. You are a great director and you were always quick in responding to PhD related questions. **Fredrik Piehl,** I am grateful for your valuable inputs in our study and for critically evaluating the manuscript**. Mohsen,** thanks for being a good lab manager and for organizing the group activities. **Brinda,** for sharing all the India-Sweden related stories during fika times and **Venus** for helping me with my ordering and travel.

Daniel Ramsköld and **Francesco,** you are the two brilliant bioinformaticians I have come across and thank you for generously sharing your time and expertise. **Sunjay** and **Ewoud** thanks for all the good time with you outside the lab. You guys will have a bright future! Thank you **Sabrina** for all help during the preparation for half time seminar and PhD dissertation.The current Epi Group **Lara, Maria, Eliane, Majid** and **Galina** and the colleagues from comped group **Gilad, Imad, Rubin, Narsis, Soudabeh** and **Angelika** and other colleagues from MS group **Jenny, Ryan** and **Kerstin**. Thank you guys for all the support throughout the years. **Roham, Harald, Xingmei** and **Marie** it was great to be a part of your projects. People in Neuroimmunology unit: **Karl, André, Rux, Susanne Neumann, Susanna Brauner, Faiez, Rasmus, Mathias, Jinming, Keying, Ramil and Hannes**. Thank you for your friendliness and chats on science and life. **Leonid, Lina** and **Klementy** I am very grateful to you, for answering all the questions related to genetics. Last not the least, I would like to thank **CMM** and **CMM-IT** for providing a great environment for my PhD studies. Thank you **Daniel Uhevag** and **Olle** for all the support and help for my data intensive projects. Thank you, **Ilgar, Alex and Husain** my bioinformatics friends and my PhD mentor **Lars Arvestad** for all the good advices.

Siby, Maria, Lince, Dany, Visakh and Arya; Thanks for your friendship and for all that exploration trips we made in Sweden and abroad.

Dad and Mom, I wouldn't have done a PhD without your support and for allowing me to make my choices in life. You always valued good education and international exposure. Now I understand what you meant during my younger age! Liz, thank you for being a wonderful partner and adjusting to my life style. You have so much of positive energy and I have benefited a lot from it. Thanks to my brothers and cousins who constantly asked about my ongoing research. It was always encouraging! Finally I am thanking God for all the support!

7 REFERENCES

1. Kumar DR, Aslinia F, Yale SH, Mazza JJ. Jean-martin charcot: The father of neurology. Clin Med Res. 2011;9: 46–49. doi:10.3121/cmr.2009.883

2. Ahlgren C, Odén A, Lycke J. High nationwide incidence of multiple sclerosis in Sweden. PLoS One. 2014;9. doi:10.1371/journal.pone.0108599

3. Dendrou CA, Fugger L, Friese MA. Immunopathology of multiple sclerosis. Nature Reviews Immunology. 2015. pp. 545–558. doi:10.1038/nri3871

4. Kearney H, Altmann DR, Samson RS, Yiannakas MC, Wheeler-Kingshott CAM, Ciccarelli O, et al. Cervical cord lesion load is associated with disability independently from atrophy in MS. Neurology. 2015; doi:10.1212/WNL.0000000000001186

5. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. The Lancet Neurology. 2018. doi:10.1016/S1474-4422(17)30470-2

6. Link H, Huang YM. Oligoclonal bands in multiple sclerosis cerebrospinal fluid: An update on methodology and clinical usefulness. Journal of Neuroimmunology. 2006. doi:10.1016/j.jneuroim.2006.07.006

7. Freedman MS, Thompson EJ, Deisenhammer F, Giovannoni G, Grimsley G, Keir G, et al. Recommended standard of cerebrospinal fluid analysis in the diagnosis of multiple sclerosis: A consensus statement. Archives of Neurology. 2005. doi:10.1001/archneur.62.6.865

8. Compston A, Coles A. Multiple sclerosis. The Lancet. 2008. doi:10.1016/S0140-6736(08)61620-7

9. Dendrou CA, Fugger L. Immunomodulation in multiple sclerosis: promises and pitfalls. Current Opinion in Immunology. 2017. doi:10.1016/j.coi.2017.08.013

10. Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nature Reviews Neurology. 2016. pp. 26–36. doi:10.1038/nrneurol.2016.187

11. Baranzini SE, Mudge J, Van Velkinburgh JC, Khankhanian P, Khrebtukova I, Miller NA, et al. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature. 2010; doi:10.1038/nature08990

12. Lopes Pinheiro MA, Kooij G, Mizee MR, Kamermans A, Enzmann G, Lyck R, et al. Immune cell trafficking across the barriers of the central nervous system in multiple sclerosis and stroke. Biochimica et Biophysica Acta - Molecular Basis of Disease. 2016. doi:10.1016/j.bbadis.2015.10.018

13. Fan X, Lin C, Han J, Jiang X, Zhu J, Jin T. Follicular Helper CD4 $^{+}$ T Cells in Human Neuroautoimmune Diseases and Their Animal Models. Mediators Inflamm. 2015; doi:10.1155/2015/638968

14. Lalive PH. Autoantibodies in inflammatory demyelinating diseases of the central nervous system. Swiss Med Wkly Off J Swiss Soc Infect Dis Swiss Soc Intern Med Swiss Soc Pneumol. 2008; doi:/aop/smw-aop12283

15. Disanto G, Morahan JM, Barnett MH, Giovannoni G, Ramagopalan S V. The evidence for a role

of B cells in multiple sclerosis. Neurology. 2012. doi:10.1212/WNL.0b013e318249f6f0

16. Cheng Y, Sun L, Xie Z, Fan X, Cao Q, Han J, et al. Diversity of immune cell types in multiple sclerosis and its animal model: Pathological and therapeutic implications. Journal of Neuroscience Research. 2017. doi:10.1002/jnr.24023

17. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015. doi:10.1038/nature15393

18. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011. doi:10.1534/genetics.110.120907

19. Zheleznyakova GY, Piket E, Marabita F, Pahlevan Kakhki M, Ewing E, Ruhrmann S, et al. Epigenetic research in multiple sclerosis: progress, challenges, and opportunities. Physiol Genomics. American Physiological Society; 2017;49: 447–461. doi:10.1152/physiolgenomics.00060.2017

20. O'Gorman C, Lin R, Stankovich J, Broadley SA. Modelling genetic susceptibility to multiple sclerosis with family data. Neuroepidemiology. 2013; doi:10.1159/000341902

21. Westerlind H, Ramanujam R, Uvehag D, Kuja-Halkola R, Boman M, Bottai M, et al. Modest familial risks for multiple sclerosis: A registry-based study of the population of Sweden. Brain. 2014;137: 770–778. doi:10.1093/brain/awt356

22. Jersild C, Svejgaard A, Fog T. HL-A antigens and multiple sclerosis. Lancet (London, England). 1972; doi:10.1038/nm.3485

23. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: Allele variant databases. Nucleic Acids Res. 2015; doi:10.1093/nar/gku1161

24. Patsopoulos NA, Barcellos LF, Hintzen RQ, Schaefer C, van Duijn CM, Noble JA, et al. Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. PLoS Genet. 2013; doi:10.1371/journal.pgen.1003926

25. Fogdell-Hahn A, Ligers A, Grønning M, Hillert J, Olerup O. Multiple sclerosis: A modifying influence of HLA class I genes in an HLA class II associated autoimmune disease. Tissue Antigens. 2000; doi:10.1034/j.1399-0039.2000.550205.x

26. Sawcer S. The complex genetics of multiple sclerosis: Pitfalls and prospects. Brain. 2008. doi:10.1093/brain/awn081

27. Hollenbach JA, Oksenberg JR. The immunogenetics of multiple sclerosis: A comprehensive review. Journal of Autoimmunity. 2015. doi:10.1016/j.jaut.2015.06.010

28. Moutsianas L, Jostins L, Beecham AH, Dilthey AT, Xifara DK, Ban M, et al. Class II HLA interactions modulate genetic risk for multiple sclerosis. Nat Genet. 2015;47: 1107–1113. doi:10.1038/ng.3395

29. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, et al. Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med. 2007; doi:10.1056/NEJMoa073493

30. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011;476: 214–219. doi:101038/nature10251

31. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Research and Therapy. 2011. doi:10.1186/ar3204

32. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet. 2013; doi:10.1038/ng.2770

33. IMSGC, Patsopoulos N, Baranzini SE, Santaniello A, Shoostari P, Cotsapas C, et al. The Multiple Sclerosis Genomic Map: Role of peripheral immune cells and resident microglia in susceptibility. bioRxiv. 2017; 143933. doi:10.1101/143933

34. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017; doi:10.1093/nar/gkw1133

35. Agarwala V, Flannick J, Sunyaev S, Altshuler D. Evaluating empirical bounds on complex disease genetic architecture. Nat Genet. 2013; doi:10.1038/ng.2804

36. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010. doi:10.1016/j.cell.2010.03.032

37. Mitrovic M, Patsopoulos N, Beecham A, Dankowski T, Goris A, Dubois B, et al. Low frequency and rare coding variation contributes to multiple sclerosis risk. bioRxiv. 2018;

38. Nishizaki SS, Boyle AP. Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. Trends in Genetics. 2017. doi:10.1016/j.tig.2016.10.008

39. Kircher M, Witten DMD, Jain P, O'Roak BJBBJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46: 310–5. doi:10.1038/ng.2892

40. Westra HJ, Franke L. From genome to function by studying eQTLs. Biochimica et Biophysica Acta - Molecular Basis of Disease. 2014. doi:10.1016/j.bbadis.2014.04.024

41. Jansen RC, Nap JP. Genetical genomics: The added value from segregation. Trends Genet. 2001; doi:10.1016/S0168-9525(01)02310-1

42. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, et al. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet. 2004; doi:10.1086/426461

43. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012; doi:10.1093/bioinformatics/bts163

44. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science. 2014;344: 519–23. doi:10.1126/science.1249547

45. Folkersen L, F van't H, E C, HE A, GK H, U H, et al. Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. Circ Cardiovasc Genet. 2010;3: 365–73.

46. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; doi:10.1038/nprot.2008.211

47. Kang EY, Martin L, Mangul S, Isvilanonda W, Zou J, Ben-David E, et al. Discovering SNPs

regulating human gene expression using allele specific expression from RNA-Seq Data. Genetics. 2016; doi:10.1534/genetics.115.177246

48. James T, Lindén M, Morikawa H, Fernandes SJ, Ruhrmann S, Huss M, et al. Impact of genetic risk loci for multiple sclerosis on expression of proximal genes in patients. Hum Mol Genet. 2018;27: 912–928.

49. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic variation in human gene expression. Science (80-). 2002; doi:10.1126/science.1072545

50. Moyerbrailean GA, Richards AL, Kurtz D, Kalita CA, Davis GO, Harvey CT, et al. High-throughput allele-specific expression across 250 environmental conditions. Genome Res. 2016; doi:10.1101/gr.209759.116

51. Guerreiro-Cacais AO, Laaksonen H, Flytzani S, N'diaye M, Olsson T, Jagodic M. Translational utility of experimental autoimmune encephalomyelitis: Recent developments. Journal of Inflammation Research. 2015. doi:10.2147/JIR.S76707

52. Baud A, Hermsen R, Guryev V, Stridh P, Graham D, McBride MW, et al. Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nat Genet. 2013; doi:10.1038/ng.2644

53. Stüve O, Youssef S, Slavin AJ, King CL, Patarroyo JC, Hirschberg DL, et al. The role of the MHC class II transactivator in class II expression and antigen presentation by astrocytes and in susceptibility to central nervous system autoimmune disease. J Immunol. 2002; doi:10.4049/jimmunol.169.12.6720

54. Gyllenberg A, Piehl F, Alfredsson L, Hillert J, Bomfim IL, Padyukov L, et al. Variability in the CIITA gene interacts with HLA in multiple sclerosis. Genes Immun. 2014; doi:10.1038/gene.2013.71

55. Laaksonen H, Guerreiro-Cacais AO, Adzemovic MZ, Parsa R, Zeitelhofer M, Jagodic M, et al. The multiple sclerosis risk gene IL22RA2 contributes to a more severe murine autoimmune neuroinflammation. Genes Immun. 2014; doi:10.1038/gene.2014.36

56. Jagodic M, Colacios C, Nohra R, Dejean AS, Beyeen AD, Khademi M, et al. A role for VAV1 in experimental autoimmune encephalomyelitis and multiple sclerosis. Sci Transl Med. 2009;1: 10ra21. doi:10.1126/scitranslmed.3000278

57. Gold R, Linington C, Lassmann H. Understanding pathogenesis and therapy of multiple sclerosis via animal models: 70 Years of merits and culprits in experimental autoimmune encephalomyelitis research. Brain. 2006. doi:10.1093/brain/awl075

58. Yednock TA, Cannon C, Fritz LC, Sanchez-Madrid F, Steinman L, Karin N. Prevention of experimental autoimmune encephalomyelitis by antibodies against alpha 4 beta 1 integrin. Nature. 1992; doi:10.1038/356063a0

59. Teitelbaum D, Meshorer a, Hirshfeld T, Arnon R, Sela M. Suppression of experimental allergic encephalomyelitis by a synthetic polypeptide. Eur J Immunol. 1971; doi:10.1002/eji.1830010406

60. Fletcher JM, Lalor SJ, Sweeney CM, Tubridy N, Mills KHG. T cells in multiple sclerosis and experimental autoimmune encephalomyelitis. Clinical and Experimental Immunology. 2010. doi:10.1111/j.1365-2249.2010.04143.x

61. Furtado GC, Marcondes MCG, Latkowski J-A, Tsai J, Wensky A, Lafaille JJ. Swift Entry of Myelin-

Specific T Lymphocytes into the Central Nervous System in Spontaneous Autoimmune Encephalomyelitis. J Immunol. 2008; doi:10.4049/jimmunol.181.7.4648

62. Rothhammer V, Heink S, Petermann F, Srivastava R, Claussen MC, Hemmer B, et al. Th17 lymphocytes traffic to the central nervous system independently of α4 integrin expression during EAE. J Exp Med. 2011; doi:10.1084/jem.20110434

63. O'Connor RA, Prendergast CT, Sabatos CA, Lau CWZ, Leech MD, Wraith DC, et al. Cutting edge: Th1 cells facilitate the entry of Th17 cells to the central nervous system during experimental autoimmune encephalomyelitis. J Immunol. 2008; doi:10.4049/jimmunol.181.6.3750

64. Bettelli E, Sullivan B, Szabo SJ, Sobel RA, Glimcher LH, Kuchroo VK. Loss of T-bet, But Not STAT1, Prevents the Development of Experimental Autoimmune Encephalomyelitis. J Exp Med. 2004; doi:10.1084/jem.20031819

65. Cua DJ, Sherlock J, Chen Y, Murphy CA, Joyce B, Seymour B, et al. Interleukin-23 rather than interleukin-12 is the critical cytokine for autoimmune inflammation of the brain. Nature. 2003; doi:10.1038/nature01355

66. Becher B, Durell BG, Noelle RJ. Experimental autoimmune encephalitis and inflammation in the absence of interleukin-12. J Clin Invest. 2002; doi:10.1172/JCI0215751

67. Rangachari M, Kuchroo VK. Using EAE to better understand principles of immune function and autoimmune pathology. Journal of Autoimmunity. 2013. doi:10.1016/j.jaut.2013.06.008

68. Fillatreau S, Sweenie CH, McGeachy MJ, Gray D, Anderton SM. B cells regulate autoimmunity by provision of IL-10. Nat Immunol. 2002; doi:10.1038/ni833

69. Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell. 2004. doi:10.1016/S0092-8674(04)00045-5

70. Baulina NM, Kulakova OG, Favorova OO. MicroRNAs: The role in autoimmune inflammation. Acta Naturae. 2016.

71. Bi Y, Liu G, Yang R. MicroRNAs: Novel regulators during the immune response. Journal of Cellular Physiology. 2009. doi:10.1002/jcp.21639

72. Ha T-Y. The Role of MicroRNAs in Regulatory T Cells and in the Immune Response. Immune Netw. 2011; doi:10.4110/in.2011.11.1.11

73. O'Connell RM, Kahn D, Gibson WSJ, Round JL, Scholz RL, Chaudhuri AA, et al. MicroRNA-155 promotes autoimmune inflammation by enhancing inflammatory T cell development. Immunity. 2010; doi:10.1016/j.immuni.2010.09.009

74. Lewkowicz P, Cwikli ska H, Mycko MP, Cichalewska M, Domowicz M, Lewkowicz N, et al. Dysregulated RNA-Induced Silencing Complex (RISC) Assembly within CNS Corresponds with Abnormal miRNA Expression during Autoimmune Demyelination. J Neurosci. 2015; doi:10.1523/JNEUROSCI.4794-14.2015

75. Schneider M V., Orchard S. Omics technologies, data and bioinformatics principles. Methods in molecular biology (Clifton, N.J.). 2011. doi:10.1007/978-1-61779-027-0_1

76. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016; doi:10.1093/bib/bbv108

77. McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491: 56–65. doi:10.1038/nature11632

78. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 2012; doi:10.1101/gr.135350.111

79. Shay T, Kang J. Immunological Genome Project and systems immunology. Trends in Immunology. 2013. doi:10.1016/j.it.2013.03.004

80. Wingender E, Dietze P, Karas H, Knüppel R. TRANSFAC: A database on transcription factors and their DNA binding sites. Nucleic Acids Research. 1996. doi:10.1093/nar/24.1.238

81. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nature Biotechnology. 2010. doi:10.1038/nbt1210-1248

82. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. BMC systems biology. 2014. doi:10.1186/1752-0509-8-S2-I1

83. LaFramboise T. Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. Nucleic Acids Research. 2009. doi:10.1093/nar/gkp552

84. Ceballos FC, Hazelhurst S, Ramsay M. Assessing runs of Homozygosity: A comparison of SNP Array and whole genome sequence low coverage data. BMC Genomics. 2018; doi:10.1186/s12864-018-4489-0

85. Distefano JK, Taverna DM. Technological issues and experimental design of gene association studies. Methods Mol Biol. 2011; doi:10.1007/978-1-61737-954-3_1

86. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol. 2012; doi:10.1371/journal.pcbi.1002822

87. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics. 2011. doi:10.1038/nrg2986

88. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. www.genome.gov/sequencingcostsdata. 2016;

89. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genet Med. 2018; doi:10.1038/gim.2017.247

90. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012;44: 623–630. doi:10.1038/ng.2303

91. Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, et al. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. RNA. 2010; doi:10.1261/rna.1947110

92. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;

doi:10.1371/journal.pone.0078644

93. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; doi:10.1038/nature08872

94. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; doi:10.1038/nature08903

95. Harrell FE. Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis. Springer Series in Statistics. 2001. doi:10.1007/978-1-4757-3462-1

96. Perez-Riverol Y, Kuhn M, Vizcaíno JA, Hitz MP, Audain E. Accurate and fast feature selection workflow for high-dimensional omics data. PLoS One. 2017; doi:10.1371/journal.pone.0189875

97. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: Mathematical aspects. BMC Bioinformatics. 2016; doi:10.1186/s12859-015-0857-9

98. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform. 2016; doi:10.1093/bib/bbw114

99. Huang S, Chaudhary K, Garmire LX. More is better: Recent progress in multi-omics data integration methods. Frontiers in Genetics. 2017. doi:10.3389/fgene.2017.00084

100. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biology. 2017. doi:10.1186/s13059-017-1215-1

101. Kular L, Liu Y, Ruhrmann S, Zheleznyakova G, Marabita F, Gomez-Cabrero D, et al. DNA methylation as a mediator of HLA-DRB1 15:01 and a protective variant in multiple sclerosis. Nat Commun. 2018;9. doi:10.1038/s41467-018-04732-5

102. Kozomara A, Griffiths-Jones S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014; doi:10.1093/nar/gkt1181

103. Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM. MiRanalyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res. 2011; doi:10.1093/nar/gkr247

104. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol. 2010; doi:10.1038/nsmb.2115

105. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. 2010; doi:10.1186/gb-2010-11-8-r90

106. Riffo-Campos ÁL, Riquelme I, Brebi-Mieville P. Tools for sequence-based miRNA target prediction: What to choose? International Journal of Molecular Sciences. 2016. doi:10.3390/ijms17121987

107. Hedström AK, Hillert J, Olsson T, Alfredsson L. Nicotine might have a protective effect in the etiology of multiple sclerosis. Mult Scler J. 2013;19: 1009–1013. doi:10.1177/1352458512471879

108. Hedström AK, Bäärnhielm M, Olsson T, Alfredsson L. Tobacco smoking, but not Swedish snuff use, increases the risk of multiple sclerosis. Neurology. 2009;73: 696–701. doi:10.1212/WNL.0b013e3181b59c40

109. Holmén C, Piehl F, Hillert J, Fogdell-Hahn A, Lundkvist M, Karlberg E, et al. A Swedish national post-marketing surveillance study of natalizumab treatment in multiple sclerosis. Mult Scler. 2011;17: 708–719. doi:10.1177/1352458510394701

110. Khademi M, Kockum I, Andersson ML, Iacobaeus E, Brundin L, Sellebjerg F, et al. Cerebrospinal fluid CXCL13 in multiple sclerosis: a suggestive prognostic marker for the disease course. Mult Scler. 2011;17: 335–343. doi:10.1177/1352458510389102

111. Fangerau T, Schimrigk S, Haupts M, Kaeder M, Ahle G, Brune N, et al. Diagnosis of multiple sclerosis: Comparison of the Poser criteria and the new McDonald criteria. Acta Neurol Scand. 2004; doi:10.1111/j.1600-0404.2004.00246.x

112. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;11: 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

113. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26: 589–595.

114. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics. 2015; doi:10.1093/bioinformatics/btu668

115. Packer JS, Maxwell EK, O'Dushlaine C, Lopez AE, Dewey FE, Chernomorsky R, et al. CLAMMS: A scalable algorithm for calling common and rare copy number variants from exome sequencing data. Bioinformatics. 2015;32: 133–135. doi:10.1093/bioinformatics/btv547

116. Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen JB, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Nephrol. 2014; doi:10.1186/1471-2105-15-125

117. Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat Sci. 2009; doi:10.1214/09-STS307

118. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: An efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics. 2016;32: 1423–1426. doi:10.1093/bioinformatics/btw079

119. Fadista J, Manning AK, Florez JC, Groop L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet. Nature Publishing Group; 2016;24: 1202–1205. doi:10.1038/ejhg.2015.269

120. Sherry ST. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; doi:10.1093/nar/29.1.308

121. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: Public

archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014; doi:10.1093/nar/gkt1113

122. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. Oxford University Press; 2010;38: e164–e164. doi:10.1093/nar/gkq603

123. Jeng XJ, Daye ZJ, Lu W, Tzeng J-Y. Rare Variants Association Analysis in Large-Scale Sequencing Studies at the Single Locus Level. PLOS Comput Biol. Public Library of Science; 2016;12: e1004993.

124. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17: 122. doi:10.1186/s13059-016-0974-4

125. Nicolae DL. Association Tests for Rare Variants. Annu Rev Genomics Hum Genet. Annual Reviews; 2016;17: 117–130. doi:10.1146/annurev-genom-083115-022609

126. Li B, Leal S. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; doi:10.1016/j.ajhg.2008.06.024.

127. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; doi:10.1016/j.ajhg.2011.05.029

128. Sun J, Zheng Y, Hsu L. A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. Genet Epidemiol. 2013; doi:10.1002/gepi.21717

129. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

130. Anders S, Pyl PT, Huber W, Camp JG, Vernot B, Köhler K, et al. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31: 166–169. doi:10.1093/bioinformatics/btu638

131. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13: 204–216. doi:10.1093/biostatistics/kxr054

132. Li J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013;22: 519–536. doi:10.1177/0962280211428386

133. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. 2014;10. doi:10.1371/journal.pgen.1004383

134. Aaron M. Newman, Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn and AA, Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2016;21: 193–201. doi:10.1016/j.molmed.2014.11.008.Mitochondria

135. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nat Commun. 2016; doi:10.1038/ncomms12817

136. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. Genome Biol. 2015; doi:10.1186/s13059-015-0762-6

137. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013. doi:10.1038/ng.2653

138. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. Mol Cell Proteomics. 2014; doi:10.1074/mcp.M113.035600

139. Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A. RNA-Seq Atlas-a reference database for gene expression profiling in normal tissue by next-generation sequencing. Bioinformatics. 2012; doi:10.1093/bioinformatics/bts084

140. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; doi:10.1038/nature07509

141. Asmann YW, Necela BM, Kalari KR, Hossain A, Baker TR, Carr JM, et al. Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. Cancer Res. 2012; doi:10.1158/0008-5472.CAN-11-3142

142. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25: 1105–11. doi:10.1093/bioinformatics/btp120

143. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; doi:10.1038/nbt.1621

144. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; doi:10.1093/bioinformatics/bts034

145. Pavlidis P. Using ANOVA for gene selection from microarray studies of the nervous system. Methods. 2003; doi:10.1016/S1046-2023(03)00157-9

146. Kockum I, Alferdsson L, Olsson T. Genetic and Environmental Risk Factors for Multiple Sclerosis—A Role for Interaction Analysis. In Between the Lines of Genetic Code, Genetic Interactions in Understanding Disease and Complex Phenotypes Edited by Padyukov L. First. San Diego CA USA, London, UK and Waltham MA USA.: Academic Press; 2014. pp. 124–133.

147. Amadio S, Parisi C, Piras E, Fabbrizio P, Apolloni S, Montilli C, et al. Modulation of P2X7 Receptor during Inflammation in Multiple Sclerosis. Front Immunol. Frontiers Media S.A.; 2017;8: 1529. doi:10.3389/fimmu.2017.01529

148. Patel RC, Sen GC. PACT, a protein activator of the interferon-induced protein kinase, PKR. EMBO J. 1998;17: 4379–4390. doi:10.1093/emboj/17.15.4379

149. Jung S-H, Yim S-H, Hu H-J, Lee KH, Lee J-H, Sheen D-H, et al. Genome-wide copy number variation analysis identifies deletion variants associated with ankylosing spondylitis. Arthritis Rheumatol (Hoboken, NJ). 2014;66: 2103–12. doi:10.1002/art.38650

150. Gu BJ, Field J, Dutertre S, Ou A, Kilpatrick TJ, Lechner-Scott J, et al. A rare P2X7 variant Arg307Gln with absent pore formation function protects against neuroinflammation in multiple sclerosis. Hum Mol Genet. 2015;24: 5644–5654. doi:10.1093/hmg/ddv278

151. Zabel BA, Agace WW, Campbell JJ, Heath HM, Parent D, Roberts AI, et al. Human G Protein – coupled Receptor GPR-9-6 / CC Chemokine Receptor 9 Is Selectively Expressed on Intestinal Chemokine – mediated Chemotaxis. Cell. 1999;190: 1241–55.

152. Papadakis KA, Prehn J, Nelson V, Cheng L, Binder SW, Ponath PD, et al. The Role of Thymus-Expressed Chemokine and Its Receptor CCR9 on Lymphocytes in the Regional Specialization of the Mucosal Immune System. J Immunol. 2000;165: 5069–5076. doi:10.4049/jimmunol.165.9.5069

153. Shenoy AR, Wellington DA, Kumar P, Kassa H, Booth CJ, Cresswell P, et al. GBP5 Promotes NLRP3 inflammasome assembly and immunity in mammals. Science (80-). 2012;336: 481–485. doi:10.1126/science.1217141

154. Gris D, Ye Z, Iocca HA, Wen H, Craven RR, Gris P, et al. NLRP3 Plays a Critical Role in the Development of Experimental Autoimmune Encephalomyelitis by Mediating Th1 and Th17 Responses. J Immunol. 2010;185: 974–981. doi:10.4049/jimmunol.0904145

155. Teer JK, Mullikin JC. Exome sequencing: The sweet spot before whole genomes. Hum Mol Genet. 2010; doi:10.1093/hmg/ddq333

156. Guo Y, Long J, He J, Li CI, Cai Q, Shu XO, et al. Exome sequencing generates high quality data in non-target regions. BMC Genomics. 2012; doi:10.1186/1471-2164-13-194

157. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science (80-). 2014;343: 1246949. doi:10.1126/science.1246949

158. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, et al. Genetics of gene expression in primary immune cells identifies cell type – specific master regulators and roles of HLA alleles. Nat Genet. 2012;44: 1–10. doi:10.1038/ng.2205

159. Lopez de Lapuente A, Feliú A, Ugidos N, Mecha M, Mena J, Astobiza I, et al. Novel Insights into the Multiple Sclerosis Risk Gene ANKRD55. J Immunol. 2016;196: 4553–65. doi:10.4049/jimmunol.1501205

160. Onengut-Gumuscu S, Chen WM, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet. 2015;47: 381–386. doi:10.1038/ng.3245

161. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010. pp. 733–739. doi:10.1038/nrg2825

162. T'Hoen PAC, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol. 2013;31: 1015–1022. doi:10.1038/nbt.2702

163. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol. 2014;32: 915–925. doi:10.1038/nbt.2972

164. Fonseca NA, Marioni J, Brazma A. RNA-Seq gene profiling - A systematic empirical comparison. PLoS One. 2014;9. doi:10.1371/journal.pone.0107026

165. Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, et al. Limited

statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat Genet. 2017; doi:10.1038/ng.3795

166. Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-). 2015; doi:10.1126/science.1262110

167. McKay KA, Kwan V, Duggan T, Tremlett H. Risk factors associated with the onset of relapsing-remitting and primary progressive multiple sclerosis: A systematic review. BioMed Research International. 2015. doi:10.1155/2015/817238

168. Friese MA, Jakobsen KB, Friis L, Etzensperger R, Craner MJ, McMahon RM, et al. Opposing effects of HLA class I molecules in tuning autoreactive CD8+ T cells in multiple sclerosis. Nat Med. 2008; doi:10.1038/nm.1881

169. Sun L, Dimitromanolakis A, Faye LL, Paterson AD, Waggott D, Bull SB. BR-squared: A practical solution to the winner's curse in genome-wide scans. Hum Genet. 2011; doi:10.1007/s00439-011-0948-2

170. Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, et al. SweGen: A whole-genome data resource of genetic variability in a cross-section of the Swedish population. Eur J Hum Genet. 2017;25: 1253–1260. doi:10.1038/ejhg.2017.130

171. Sundal C, Baker M, Karrenbauer V, Gustavsen M, Bedri S, Glaser A, et al. Hereditary diffuse leukoencephalopathy with spheroids with phenotype of primary progressive multiple sclerosis. Eur J Neurol. 2015;22: 328–333. doi:10.1111/ene.12572

172. Wang Z, Sadovnick AD, Traboulsee ALL, Ross JPP, Bernales CQQ, Encarnacion M, et al. Nuclear Receptor NR1H3 in Familial Multiple Sclerosis. Neuron. 2016;90: 948–954. doi:10.1016/j.neuron.2016.04.039

173. International Multiple Sclerosis Genetics Consortium. NR1H3 p.Arg415Gln Is Not Associated to Multiple Sclerosis Risk. Neuron. 2016;92: 333–335. doi:10.1016/j.neuron.2016.09.052

174. Sriram S, Steiner I. Experimental allergic encephalomyelitis: A misleading model of multiple sclerosis. Annals of Neurology. 2005. doi:10.1002/ana.20743

175. Constantinescu CS, Farooqi N, O'Brien K, Gran B. Experimental autoimmune encephalomyelitis (EAE) as a model for multiple sclerosis (MS). British Journal of Pharmacology. 2011. doi:10.1111/j.1476-5381.2011.01302.x

176. Wang Q, Armenia J, Zhang C, Penson A V, Reznik E, Zhang L, et al. Enabling cross-study analysis of RNA-Sequencing data. bioRxiv. 2017; doi:10.1101/110734

177. Uhlen M, Hallstro m BM, Lindskog C, Mardinoglu A, Ponten F, Nielsen J. Transcriptomics resources of human tissues and organs. Mol Syst Biol. 2016; doi:10.15252/msb.20155865